Andrej Aderhold, Dirk Husmeier* and Marco Grzegorczyk

Statistical inference of regulatory networks for circadian regulation

Abstract: We assess the accuracy of various state-of-the-art statistics and machine learning methods for reconstructing gene and protein regulatory networks in the context of circadian regulation. Our study draws on the increasing availability of gene expression and protein concentration time series for key circadian clock components in *Arabidopsis thaliana*. In addition, gene expression and protein concentration time series are simulated from a recently published regulatory network of the circadian clock in *A. thaliana*, in which protein and gene interactions are described by a Markov jump process based on Michaelis-Menten kinetics. We closely follow recent experimental protocols, including the entrainment of seedlings to different light-dark cycles and the knock-out of various key regulatory genes. Our study provides relative network reconstruction accuracy scores for a critical comparative performance evaluation, and sheds light on a series of highly relevant questions: it quantifies the influence of systematically missing values related to unknown protein concentrations and mRNA transcription rates, it investigates the dependence of the performance on the network topology and the degree of recurrency, it provides deeper insight into when and why non-linear methods fail to outperform linear ones, it offers improved guidelines on parameter settings in different inference procedures, and it suggests new hypotheses about the structure of the central circadian gene regulatory network in *A. thaliana*.

Keywords: regulatory network inference; circadian clock; hierarchical Bayesian models; comparative method evaluation; ANOVA.

DOI 10.1515/sagmb-2013-0051

1 Introduction

Plants have to carefully manage their resources. The process of photosynthesis allows them to utilize sunlight to produce essential carbohydrates during the day. However, the earth's rotation predictably removes sunlight, and hence the opportunity for photosynthesis, for a significant part of each day, and plants need to orchestrate the accumulation, utilization and storage of photosynthetic products in the form of starch over the daily cycle to avoid periods of starvation, and thus optimize growth rates.

In the last few years, substantial progress has been made to model the central processes of circadian regulation, i.e., the mechanism of internal time-keeping that allows the plant to anticipate each new day, at the molecular level (Guerriero et al., 2012; Pokhilko et al., 2012). Moreover, simple mechanistic models have been developed to describe the feedback between carbon metabolism and the circadian clock, by which the plant adjusts the rates of starch accumulation and consumption in response to changes in the light-dark cycle (Feugier and Satake, 2012). What is needed is the elucidation of the detailed structure of the molecular regulatory networks and signaling pathways of these processes, by utilization and integration of transcriptomic, proteomic and metabolic concentration profiles that become increasingly available from international

***Corresponding author: Dirk Husmeier,** School of Mathematics and Statistics, University of Glasgow, 15 University Gardens, Glasgow G12 8QW, UK, e-mail: dirk.husmeier@glasgow.ac.uk
Andrej Aderhold: School of Mathematics and Statistics, University of Glasgow, 15 University Gardens, Glasgow G12 8QW, UK; and School of Biology, Sir Harold Mitchell Building, University of St Andrews, St Andrews, Fife KY16 9TH, UK
Marco Grzegorczyk: Johann Bernoulli Institute (JBI), Groningen University, Nijenborgh 9, 9747 AG Groningen, The Netherlands

research collaborations like AGRON-OMICS[1] and TiMet (2014). The inference of molecular regulatory networks from post-genomic data has been a central topic in computational systems biology for over a decade. Following up on the seminal paper of Friedman et al. (2000), a variety of methods have been proposed and several procedures have been pursued to objectively assess the network reconstruction accuracy (Husmeier, 2003; Werhli et al., 2006; Weirauch et al., 2013). The objective of the present article is to complement these studies in six important respects. Firstly, we have taken a particular focus on circadian regulation. To this end, we have taken the central circadian clock network in *Arabidopsis thaliana*, as published by Guerriero et al. (2012), as a ground truth for evaluation, and closely followed recent experimental protocols for data generation, including the entrainment of seedlings to different light-dark cycles, and the knock-out of various key regulatory genes. To make the data generated from this network as realistic as possible, we have modeled gene and protein interactions as a Markov jump process (MJP) based on Michaelis-Menten kinetics. This is to be preferred over mechanistic models based on ordinary differential equations (used e.g., by Pokhilko et al., 2012), as MJPs capture the intrinsic stochasticity of molecular interactions. MJPs also avoid artefacts that result from the dynamics of ordinary differential equations converging to stable limit cycles with completely regular oscillations, which are never observed in actual experiments (Guerriero et al., 2012). Secondly, we have assessed the impact of missing values on the reconstruction task. Protein-gene interactions affect transcription rates, but both these rates as well as protein concentrations might not be available from the wet lab assays. In such situations, mRNA concentrations have to be taken as proxy for protein concentrations, and rates have to be approximated by finite difference quotients. For both approximations, we have quantified the ensuing deterioration in network reconstruction accuracy. Thirdly, we have investigated the dependence of the network reconstruction accuracy on the degree of connectivity and recurrency in the network topology. The central circadian clock network is densely connected with several tight feedback loops. However, we expect the regulatory network, via which the clock acts on carbon metabolism, to be sparser and with more feed-forward structures. In our study we have therefore quantified how the network reconstruction depends on the degree of recurrency, and how the performance varies as critical feedback cycles are pruned. Fourthly, we have investigated the effect of non-linear transformations of the data, suggested by the underlying chemical kinetic equations (Michaelis-Menten), and we have proposed a novel combination of hierarchical Bayesian models with multiple change point processes. Fifthly, we have included a substantial range of different state-of-the-art models, which to our knowledge has not been attempted before. This includes mutual information based methods, graphical Gaussian models, sparse regression methods, automatic relevance determination, hierarchical Bayesian regression, change-point processes, Gaussian mixture models, Bayesian networks, Bayesian spline autoregression models, state space models, and Gaussian processes. We have carried out a systematic comparative model evaluation with an ANOVA scheme to distinguish genuine differences in model performance from exogenous factors and confounding effects. Finally, our study includes a performance evaluation on novel qRT-PCR gene expression time series from *A. thaliana*, which was provided by the TiMet project (TiMet, 2014).

2 Method overview

The starting point of our study is the mathematical formulation of transcriptional regulation introduced by Barenco et al. (2006),

$$y_{g,t} = \frac{dx_{g,t}}{dt} = \alpha_g + f_g(\mathbf{x}_{\mathbf{r}_g,t}) - \lambda_g x_{g,t} \tag{1}$$

where $x_{g,t}$ is the mRNA concentration of gene g at time t, α_g is the basal transcription rate for gene g, λ_g is the mRNA degradation rate for gene g, $f_g(\cdot)$ is an unknown regulation function, and $\mathbf{x}_{\mathbf{r}_g,t}$ is the set of gene expression

1 https://www.agronomics.ethz.ch/.

values of the putative regulators $\boldsymbol{\pi}_g$ of gene g at time t, as explained above. This fundamental equation provides the basis for learning and inference in systems biology, as e.g., described by Lawrence et al. (2010). A common approach is to approximate the time derivative on the left-hand side by a finite difference quotient:

$$\frac{dx_{g,t}}{dt} \approx \frac{x_{g,t+\Delta t} - x_{g,t}}{\Delta t} \tag{2}$$

which for a unit time delay $\Delta t=1$ leads to

$$x_{g,t+1} = x_{g,t} + \alpha_g + f_g(\mathbf{x}_{\boldsymbol{\pi}_g}, t) - \lambda_g x_{g,t} = h(x_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t}) \tag{3}$$

for some function $h(x_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t})$. This equation provides the basis for a variety of "dynamic" algorithms, including dynamic Bayesian networks (Husmeier, 2003), time-delay mutual information methods (Zoppoli et al., 2010) and time-shifted regression methods (Morrissey et al., 2011). However, as we demonstrate in more detail in Section 5.5, the finite difference approximation of equation (2) is not particularly good, and we therefore work with the explicit representation of equation (1). This might look like a "static" method, as no time-shift operation is needed, but the dynamics are explicitly represented by the time derivative $y_{g,t} = \frac{dx_{g,t}}{dt}$. The data for our study consists of mRNA concentration profiles $\{x_{1,t},\ldots,x_{N,t}\}_{t=1,\ldots,T}$ and associated protein concentration profiles $\{x_{p,1,t},\ldots,x_{p,N^*,t}\}_{t=1,\ldots,T}$. For each individual gene $g=1,\ldots,N$ we use its observed concentrations $x_{g,1},\ldots,x_{g,T}$ to compute the gradients $y_{g,t}$ $(t=1,\ldots,T)$, and we then consider the gradients $y_{g,1},\ldots,y_{g,T}$ as realizations of a target variable y_g, which monitors the transcription rates of gene g over time. Henceforth, we refer to the variable y_g and its realizations $y_{g,t}$ $(t=1,\ldots,T)$ as the mRNA concentration time derivatives, gradients, or transcription rates synonymously. Mathematically, our goal is to find the regulators of each target variable y_g $(g=1,\ldots,N)$, i.e., to identify variables with an effect on the transcription rates y_g of gene g. We distinguish two scenarios: In the *incomplete* data scenario the potential regulators for target variable y_g are the observed mRNA concentrations of the genes $\{x_{1,t},\ldots,x_{N,t}\}_{t=1,\ldots,T}$, including the concentrations $\{x_{g,t}\}_{t=1,\ldots,T}$ of the target gene g themselves. In the *complete* data scenario we consider the protein rather than the mRNA concentrations as potential regulators. To be consistent with the fundamental equation of transcription, equation (1), $x_{g,t}$ will always be included in either scenario; we won't mention that explicitly in the text. More formally, we introduce the following symbols and notation.

2.1 Notation

For the models that we will use to infer the network interactions, we have target variables y_g $(g=1,\ldots,N)$, each representing the mRNA time derivative or gradient of a particular gene g. The realizations of each target variable y_g can then be written as a vector $\mathbf{y}_g=(y_{g,1},\ldots,y_{g,T})^\top$ where $y_{g,t}$ is the realization (or observation) of y_g at data point t. As we consider sets of time series we refer to the index t as time point and data point synonymously, in particular we also say that $y_{g,t}$ is the observation of y_g at time t. For each gene g there are G_g *potential* regulators, $x_1^g,\ldots,x_{G_g}^g$, which are either gene or protein concentrations.[2] The task is to infer a set of regulators $\boldsymbol{\pi}_g$ with $\boldsymbol{\pi}_g \subset \{x_1^g,\ldots,x_{G_g}^g\}$ for each target variable y_g.[3] The collection of regulators $\{\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_N\}$ can then be thought of as a regulatory interaction graph, $\mathcal{M}$. In $\mathcal{M}$ the regulators and the target variables represent the nodes and from each regulator in $\boldsymbol{\pi}_g$ a directed edge is pointing to the target node y_g. Hence, in terms of graphical models the graph $\mathcal{M}$ possesses a bipartite structure, where the *potential* regulators $x_1^g,\ldots,x_{G_g}^g$ are the potential parent nodes of the target variable y_g $(g=1,\ldots,N)$, and there is a directed edge from x_i^g to y_g

2 Note that the sets of potential regulators are defined for each gene g specifically. That is, the potential regulators for two target variables y_g and $y_{g'}$ can be different, e.g., if certain (biologically-motivated) restrictions are imposed.

3 For consistency with the fundamental equation of transcription, equation (1), we will enforce that each regulator set $\boldsymbol{\pi}_g$ for y_g contains the concentration x_g of g, symbolically $x_g \in \boldsymbol{\pi}_g$.

in $\mathcal{M}$, symbolically $x_i^g \to y_g$, if $x_i^g \in \boldsymbol{\pi}_g$. In regression models the regulators are usually referred to as covariates, and throughout the paper we therefore use the terms regulator(s), parent node(s) and covariate(s) interchangeably.

In regression models the observations of all the *potential* covariates of the target y_g can be collected in a design matrix $\mathbf{X}_g$ such that each row of $\mathbf{X}_g$ corresponds to a covariate and contains all T observations of that particular covariate. An additional row with constant elements equal to 1 is added to $\mathbf{X}_g$ to take the intercept into account. In addition, for a fixed subset of covariates, $\boldsymbol{\pi}_g$, we define $\mathbf{X}_{\pi_g}$ to be the sub-matrix of the full design matrix, $\mathbf{X}_g$, where all rows that belong to covariates which are not in $\boldsymbol{\pi}_g$ have been deleted. To paraphrase that, in the restricted design matrix $\mathbf{X}_{\pi_g}$ we keep only those rows of $\mathbf{X}_g$ that correspond either to the intercept or to the covariates in the set $\boldsymbol{\pi}_g$.

For non-regression models we additionally define two vectors. For $t=1,\ldots,T$ let $\mathbf{x}_{g,t} := (x_{1,t}^g,\ldots,x_{G_g,t}^g)^\top$ denote the vector of the concentrations of all G_g *potential* regulators for gene g at time t. Let $\mathbf{z}_{g,t} := (y_{g,t}, \mathbf{x}_{g,t}^\top)^\top$ extend the vector $\mathbf{x}_{g,t}$ by including the value of the response $y_{g,t}$, i.e., the derivative of the concentration of the target gene g at time t ($t=1,\ldots,T$). In addition, for a fixed subset of regulators, $\boldsymbol{\pi}_g$, we define $\mathbf{x}_{\pi_g,t}$ and $\mathbf{z}_{\pi_g,t}$ to be the corresponding sub-vectors of $\mathbf{x}_{g,t}$ and $\mathbf{z}_{g,t}$, respectively, where all elements that do not correspond to regulators in $\boldsymbol{\pi}_g$ have been deleted.

Finally, denote by $\mathbf{X}_g^*$ and $\mathbf{X}_{\boldsymbol{\pi}_g}^*$ the sub-matrices of the design matrices $\mathbf{X}_g$ and $\mathbf{X}_{\pi_g}$ in which the constant row for the intercept has been removed. For the state-space models (SSMs), described in Section 2.9, we define $\mathbf{x}_{\cdot,t}$ as the vector of the observations of *all* potential regulators at time t.[4] A complete overview of the notation is given in Table 1.

Table 1 Overview of all symbols, introduced in Section 2.

Symbol	Short verbal description
g	target gene g ($g=1,\ldots,N$)
x_g	variable measuring the mRNA concentration of target gene g
$x_{g,t}$	variable x_g at time t ($t=1,\ldots,T$)
y_g	target (response) variable, gradient corresponding to target gene g
$y_{g,t}$	target (response) variable y_g at time t, derivative of x_g at time t ($t=1,\ldots,T$)
$\mathbf{y}_{\cdot,t}$	vector of all target variables (gradients) at time t $\mathbf{y}_{\cdot,t} := (y_{1,t},\ldots,y_{N,t})^\top$
$\mathbf{y}_g$	vector of all T observations for the target gene y_g $\mathbf{y}_g := (y_{g,1},\ldots,y_{g,T})^\top$
G_g	the number of potential regulators for target gene g
x_i^g	the i-th regulator for target gene g ($g=1,\ldots,G_g$)
$x_{i,t}^g$	the observation for the i-th regulator for target gene g at time t
$\boldsymbol{\pi}_g$	concrete set of regulators (covariates, parent nodes) for target gene g $\boldsymbol{\pi}_g \subset \{x_1^g,\ldots,x_{G_g}^g\}$
$\mathcal{M}$	the bipartite graph structure $\mathcal{M}=\{\boldsymbol{\pi}_g,\ldots,\boldsymbol{\pi}_N\}$
$\mathbf{X}_g$	full design matrix for gene g including all G_g potential regulators for g
$\mathbf{X}_{\boldsymbol{\pi}_g}$	restricted design matrix for gene g, $\mathbf{X}_g$ restricted to regulators in the set $\boldsymbol{\pi}_g$
$\mathbf{x}_{g,t}$	vector of observations for all G_g regulators of gene g at time t $\mathbf{x}_{g,t} := (x_{1,t}^g,\ldots,x_{G_g,t}^g)^\top$
$\mathbf{z}_{g,t}$	response variable for gene g and concentrations of all its G potential regulators at time t, $\mathbf{z}_{g,t} := (y_{g,t}, \mathbf{x}_{g,t}^\top)^\top$
$\mathbf{x}_{\boldsymbol{\pi}_g,t}$	vector of observations for the $\lvert\boldsymbol{\pi}_g\rvert$ regulators in $\boldsymbol{\pi}_g$ at time t
$\mathbf{z}_{\boldsymbol{\pi}_g,t}$	response variable for gene g and concentrations of its regulators in the set $\boldsymbol{\pi}_g$ at time t, $\mathbf{z}_{\boldsymbol{\pi}_g,t} := (y_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t}^\top)^\top$
$\mathbf{X}_g^*$	the matrix (or set) of all T observations for the G_g potential regulators of g similar to the full design matrix $\mathbf{X}_g$, but without the row for the intercept
$\mathbf{X}_{\boldsymbol{\pi}_g}^*$	the matrix (or set) of all T observations for the regulators in $\boldsymbol{\pi}_g$ similar to the restricted design matrix $\mathbf{X}_{\boldsymbol{\pi}_g}$, but without the intercept row
$\mathbf{x}_{\cdot,t}$	vector of observations of *all* potential regulators at time t, i.e. this vector includes every available regulator, and it is not target-specific

These notations are used throughout the paper. For more detailed descriptions see main text in Section 2.

4 Note that vector $\mathbf{x}_{\cdot,t}$ includes every available regulator without any dependency on the target gene g.

2.2 Graphical Gaussian models (GGM)

The method of graphical Gaussian models (GGMs) is based on the insight that for random vectors $\mathbf{z}$ from a multivariate Gaussian distribution, $\mathbf{z}\sim\mathcal{N}(\mathbf{0}, \mathbf{C})$, the components z_g and $z_{g'}$, corresponding e.g., to two genes g and g', are stochastically independent conditional on the remaining system

$$p(z_g, z_{g'}|\{z_i\}_{i\neq g,g'}) = p(z_g|\{z_i\}_{i\neq g,g'})\, p(z_{g'}|\{z_i\}_{i\neq g,g'}) \tag{4}$$

if and only if the corresponding element (g, g') in the inverse covariance matrix $\mathbf{C}^{-1}$ is zero. Hence, if $\mathbf{C}$ is known, an undirected graph of gene dependence structures can be obtained by connecting all genes (g, g') with $[\mathbf{C}^{-1}]_{g,g'}\neq 0$ by an (undirected) edge. In practice, $\mathbf{C}$ is unknown and has to be approximated by the empirical covariance matrix

$$\mathbf{S} = \frac{1}{(T-1)}\sum_{t=1}^{T}(\mathbf{z}_t-\bar{\mathbf{z}})(\mathbf{z}_t-\bar{\mathbf{z}})^{\top} \tag{5}$$

where $\mathbf{z}_1,\ldots,\mathbf{z}_T$ is an i.i.d. sample. If T is less than the dimension of $\mathbf{z}_t$, then the estimated covariance matrix $\mathbf{S}$ is rank deficient. To deal with this problem, two main approaches have been proposed. The first approach, proposed in Schäfer and Strimmer (2005), is to use shrinkage and replace the empirical covariance matrix $\mathbf{S}$ by the following regularized matrix:

$$\mathbf{S}^{*} = (1-\lambda)\mathbf{S} + \lambda\mathbf{I} \tag{6}$$

where $\mathbf{I}$ is the identity matrix (various alternatives are discussed by Schäfer and Strimmer, 2005) and $\lambda>0$ is a regularization parameter, which can be optimized with empirical risk minimization; see equations (8) and (10) in Schäfer and Strimmer (2005) for explicit expressions. The second approach, proposed by Friedman et al. (2008) and termed "Glasso" (for "Graphical Lasso") is to maximize the penalized likelihood subject to an L1-regularization term applied to the matrix elements:

$$\log\det(\boldsymbol{\Theta}) - \mathrm{trace}(\mathbf{S}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1 \tag{7}$$

where $\boldsymbol{\Theta}$ is the estimated inverse covariance matrix. To apply GGMs to the reconstruction of gene regulatory networks, we consider the random vectors $\mathbf{z}_{g,t} := (y_{g,t}, \mathbf{x}_{g,t}^{\top})^{\top}$, where $y_{g,t}$ is the time derivative of the mRNA concentration of target gene g at time t, see equation (1), and $\mathbf{x}_{g,t}$ is the vector of the concentrations of the G_g potential regulators at time t (t=1,…,T). For each potential target variable y_g (g=1,…,N) we extract a GGM from the sample $\{\mathbf{z}_{g,t}\}_{t=1\ldots,T}$. We then consider the first row (or column) of the resulting precision matrix. By standarization we obtain the partial correlations $\rho(y_g, x_j^g|\{x_i^g\}_{i\neq j})$ between the target variable y_g and its potential regulators x_j^g (j=1,…,G_g). Note that since the direction of causality is always directed towards the target variable y_g, the edges in the reconstructed graphs are directed, symbolically: $\{x_1^g,\ldots,x_{G_g}^g\}\rightarrow y_g$. Hence, the application of an algorithm to convert undirected into directed edges, as proposed by Opgen-Rhein and Strimmer (2007), becomes obsolete.[5] The absolute values of the partial correlations $|\rho(y_g, x_j^g|\{x_i^g\}_{i\neq j})|$ can be used to score the regulatory interactions $x_j^g\rightarrow y_g$ (g=1,…,N and j=1,…,G_g) with respect to their strengths.

2.3 Sparse regression (Lasso and Elastic Net)

An efficient and widely applied linear regression method that provides network sparsity is the least absolute shrinkage and selection operator (Lasso) introduced by Tibshirani (1995). The Lasso optimizes the regression parameters $\mathbf{w}_g$ of a linear model based on the residual sum of squares subject to an $L1$-norm penalty term,

5 Note that the repeated bi-partitioning of the genes into targets and putative regulators renders Glasso equivalent to Lasso, as discussed on page 4 of Friedman et al. (2008). Lasso will be discussed in Section 2.3.

$\lambda_1\|\mathbf{w}_g\|_1$, where λ_1 is a regularization parameter, and $\|\mathbf{w}_g\|_1$ is the sum of the absolute values of the components of $\mathbf{w}_g$:

$$\hat{\mathbf{w}}_g = \text{argmin}\{||\mathbf{y}_g - \mathbf{X}_g^\top \mathbf{w}_g||_2^2 + \lambda_1 ||\mathbf{w}_g||_1\} \tag{8}$$

For definitions of the full design matrix $\mathbf{X}_g$ and the target gradient vector $\mathbf{y}_g$ see Table 1. Equation (8) is a convex optimization problem, for which a variety of fast and effective algorithms exist (e.g., Hastie et al., 2001). The effect of equation (8) is to simultaneously shrink and prune the parameters in $\mathbf{w}_g$, thereby promoting a sparse network. The degree of sparsity depends on the regularization parameter λ_1, which can be optimized with cross-validation or information criteria, like BIC.

The shortcomings are that the Lasso will only select one predictor from a set of highly correlated variables, and that it can maximally select T variables, thereby potentially suffering from saturation effects. These difficulties are addressed with the Elastic Net method, proposed by Zou and Hastie (2005), which combines the Lasso penalty with a ridge regression penalty that constitutes a squared $L2$-norm $||\mathbf{w}_g||_2^2$:

$$\hat{\mathbf{w}}_g = \text{argmin}\{||\mathbf{y}_g - \mathbf{X}_g^\top \mathbf{w}_g||_2^2 + \lambda_1 ||\mathbf{w}_g||_1 + \lambda_2 ||\mathbf{w}_g||_2^2\} \tag{9}$$

Again, this is a convex optimization problem for which effective algorithms exist, and the regularization parameters λ_1 and λ_2 can be optimized with cross-validation or BIC. For these two approaches (Lasso and Elastic Net) we use the absolute values of the elements of the estimated regression parameter vectors $\hat{\mathbf{w}}_g$ to score the regulatory effects on y_g (g=1,…,G) with respect to their strengths.

2.4 Time-varying sparse regression (Tesla)

Ahmed and Xing (2009) proposed a time-varying generalization of sparse regression, which they called Tesla. The idea is to divide a time series into segments and perform sparse regression for each time series segment separately, subject to an additional sparsity constraint that penalizes differences between regression parameters associated with adjacent time series segments. Consider a time series of expression values for gene g, which is divided into $\mathcal{K}_g$ disjunct segments, marked by $\mathcal{K}_g+1$ demarcation points $1=\tau_{g,1}\leq\ldots\leq\tau_{g,h}\leq\ldots\leq\tau_{(\mathcal{K}_g+1)}=T$. Each segment is associated with a different set of regression parameters, $\mathbf{w}_{g,h}$, where $h\in\{1,\ldots,\mathcal{K}_g\}$ is a label that identifies the segment. To prevent over-complexity and avoid overfitting, an additional L1-norm penalty is imposed on the parameter differences for adjacent time series segments, i.e., $\mathbf{w}_{g,h}-\mathbf{w}_{g,h-1}$ for h>1:

$$\hat{\mathbf{w}}_{g,1},\ldots,\hat{\mathbf{w}}_{g,\mathcal{K}_g} = \text{argmin}\left\{\sum_{h=1}^{\mathcal{K}_g} ||\mathbf{y}_{g,h} - \mathbf{X}_{g,h}^\top \mathbf{w}_{g,h}||_2^2 + \lambda_1 \sum_{h=1}^{\mathcal{K}_g} ||\mathbf{w}_g||_1 + \lambda_2 \sum_{h=2}^{\mathcal{K}_g} ||\mathbf{w}_{g,h} - \mathbf{w}_{g,h-1}||_1\right\} \tag{10}$$

where $\mathbf{y}_{g,h}=(y_{g,(\tau_{g,h}+1)},\ldots,y_{g,\tau_{g,h+1}})^\top$ is the subvector of observations in the temporal segment h, and $\mathbf{X}_{g,h}$ is the corresponding segment specific design matrix. Given the regularization parameters λ_1 and λ_2, the optimal regression parameters $\{\hat{\mathbf{w}}_{g,h}\}$ can be found with convex programming (Ahmed and Xing, 2009). The regularization parameters themselves can be optimized with cross-validation or information criteria, like BIC. Note that different genes g can have different time series segmentations, with different values of $\mathcal{K}_g$, and that the segmentations have to be defined in advance. General guidelines for the choice of coarseness of segmentation can be found in the publication of Ahmed and Xing (2009). In our applications the segmentation is naturally suggested by the light phase, as we describe in more detail in Section 4.6.2. Also note that the original formulation of Tesla, proposed by Ahmed and Xing (2009), is for logistic regression and binary data. The modification to linear regression, as in equation (10), is straightforward and more appropriate for our application to non-binary data.

2.5 Hierarchical Bayesian regression models (HBR)

In the hierarchical Bayesian regression (HBR) approach we assume a linear regression model for the target vectors $\mathbf{y}_g$:

$$\mathbf{y}_g|(\mathbf{w}_g, \sigma_g, \mathbf{X}_{\boldsymbol{\pi}_g}) \sim \mathcal{N}(\mathbf{X}_{\boldsymbol{\pi}_g}^\top \mathbf{w}_g, \sigma_g^2 \mathbf{I}) \tag{11}$$

where $\mathbf{w}_g$ is the vector of regression parameters, $\mathbf{X}_{\boldsymbol{\pi}_g}$ is the restricted design matrix whose rows correspond to the variables in the covariate set $\boldsymbol{\pi}_g$ with an additional constant row for the intercept, and σ_g^2 is the noise variance. We impose a Gaussian prior on the regression parameter vector:

$$\mathbf{w}_g|(\sigma_g, \delta_g, \mathbf{X}_{\boldsymbol{\pi}_g}) \sim \mathcal{N}(\mathbf{0}, \delta_g, \sigma_g^2 \mathbf{I}) \tag{12}$$

The hyperparameter δ_g can be interpreted as the "signal-to-noise" (SNR) ratio (Grzegorczyk and Husmeier, 2012). For the posterior distribution we get (e.g., Bishop, 2006, Section 3.3):

$$\mathbf{w}_g|(\sigma_g, \delta_g, \mathbf{X}_{\boldsymbol{\pi}_g}, \mathbf{y}_g) \sim \mathcal{N}(\Sigma_g \mathbf{X}_{\boldsymbol{\pi}_g} \mathbf{y}_g, \sigma_g^2 \Sigma_g) \tag{13}$$

where $\Sigma_g^{-1} = \delta_g^{-1}\mathbf{I} + \mathbf{X}_{\boldsymbol{\pi}_g}\mathbf{X}_{\boldsymbol{\pi}_g}^\top$. The marginal likelihood, $p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \delta_g)$, can be obtained by application of standard results for Gaussian integrals (e.g., Bishop, 2006, Appendix B):

$$p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \delta_g) = \int p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \mathbf{w}_g)\, p(\mathbf{w}_g|\sigma_g^2, \delta_g, \mathbf{X}_{\boldsymbol{\pi}_g})\, d\mathbf{w}_g = \mathcal{N}(\mathbf{y}_g|\mathbf{0}, \sigma_g^2(\mathbf{I} + \delta_g \mathbf{X}_{\boldsymbol{\pi}_g}\mathbf{X}_{\boldsymbol{\pi}_g}^\top)) \tag{14}$$

For σ_g^{-2} and δ_g^{-2} we choose conjugate gamma priors, $\sigma_g^{-2} \sim \text{Gam}(\nu, \nu)$, and $\delta_g^{-1} \sim \text{Gam}(A_\delta, B_\delta)$.[6] The integral resulting from the marginalization over σ_g^{-2},

$$p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \delta_g) = \int_0^\infty p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \delta_g)\, p(\sigma_g^{-2}|\nu)\, d\sigma_g^{-2} \tag{15}$$

is a multivariate Student t-distribution with a closed-from solution (e.g., Bishop, 2006; Grzegorczyk and Husmeier, 2012).

Given the data for all the potential regulators of y_g, i.e., given the full design matrix $\mathbf{X}_g$, the objective is to infer the set of covariates $\boldsymbol{\pi}_g$ from the marginal posterior distribution:

$$P(\boldsymbol{\pi}_g|\mathbf{X}_g, \mathbf{y}_g, \delta_g) = \frac{P(\boldsymbol{\pi}_g)\, p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \delta_g)}{\sum_{\boldsymbol{\pi}_g^\star} P(\boldsymbol{\pi}_g^\star)\, p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g^\star}, \delta_g)} \propto P(\boldsymbol{\pi}_g)\, p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \delta_g) \tag{16}$$

where the sum in the denominator is over all valid covariate sets, $\boldsymbol{\pi}_g^\star$, $P(\boldsymbol{\pi}_g)$ is a uniform distribution over all covariate sets subject to a maximal cardinality, typically $|\boldsymbol{\pi}_g| \leq 3$. We sample sets of covariates (or regulators) $\boldsymbol{\pi}_g$, signal-to-noise hyperparameters δ_g, and noise variances σ_g^2 from the joint posterior distribution with Markov chain Monte Carlo (MCMC), following the Metropolis-Hastings within partially collapsed Gibbs scheme from Grzegorczyk and Husmeier (2012). Within that scheme, we sample covariate sets $\boldsymbol{\pi}_g$ from equation (16) with Metropolis-Hastings, using the proposal mechanism from Grzegorczyk and Husmeier (2012): given the current covariate set $\boldsymbol{\pi}_g$, randomly propose a new covariate set from the system of all covariate sets such that it can be reached (i) either by removing a single covariate from $\boldsymbol{\pi}_g$, (ii) or by adding a single covariate to $\boldsymbol{\pi}_g$, (iii) or by a covariate flip move. The (hyper-)parameters δ_g^{-1}, $\mathbf{w}_g$, and σ_g^{-2} can be sampled with Gibbs sampling steps. As shown in Grzegorczyk and Husmeier (2012), the full conditional distributions of δ_g^{-1} and $\mathbf{w}_g$ are given by:

6 We set: ν=0.005, A_δ=2, and B_δ=0.2, as in Grzegorczyk and Husmeier (2012).

$$\delta_g^{-1}|(\mathbf{w}_g, \sigma_g^2) \sim \mathrm{Gam}\left(A_\delta + \frac{|\boldsymbol{\pi}_g|+1}{2},\ B_\delta + \frac{1}{2\sigma_g^2}\mathbf{w}_g^\top \mathbf{w}_g\right) \tag{17}$$

$$\mathbf{w}_g|(\mathbf{y}_g, \mathbf{X}_{\boldsymbol{\pi}_g}, \sigma_g^2, \delta_g) \sim \mathcal{N}(\boldsymbol{\Sigma}_g^* \mathbf{X}_{\boldsymbol{\pi}_g} \mathbf{y}_g, \sigma_g^2 \boldsymbol{\Sigma}_g^*) \tag{18}$$

where $|\boldsymbol{\pi}_g|$ is the cardinality of the parent set, $\boldsymbol{\pi}_g$, and $\boldsymbol{\Sigma}_g^* = (\delta_g^{-1}\mathbf{I} + \mathbf{X}_{\boldsymbol{\pi}_g}\mathbf{X}_{\boldsymbol{\pi}_g}^\top)^{-1}$. The inverse variance hyperparameters, σ_g^{-2} can be sampled with a collapsed Gibbs sampling step, in which the regression parameter vectors, $\mathbf{w}_g$, have been integrated out. This marginalization yields (e.g., Grzegorczyk and Husmeier, 2012):

$$\sigma_g^{-2}|(\mathbf{y}_g, \mathbf{X}_{\boldsymbol{\pi}_g}, \delta_g) \sim \mathrm{Gam}\left(\nu + \frac{T}{2},\ \nu + \frac{\mathbf{y}_g^\top (\mathbf{I} + \delta_g \mathbf{X}_{\boldsymbol{\pi}_g}^\top \mathbf{X}_{\boldsymbol{\pi}_g})^{-1}\mathbf{y}_g}{2}\right) \tag{19}$$

where T is the number of observations. A compact representation of the relationships among the (hyper-) parameters of the Bayesian regression model is given in the left panel of Figure 1.

A straightforward extension of the HBR method is to include non-linear terms in the design matrix $\mathbf{X}_{\boldsymbol{\pi}_g}$. In our study we tested, as an alternative to the HBR model just described, the inclusion of quadratic and inverse terms. So for a set of regulators $\boldsymbol{\pi}_g = \{A, B\}$, the columns of $\mathbf{X}_{\boldsymbol{\pi}_g}$, $[1, x_A(t), x_B(t)]'$ are replaced by $[1, x_A(t), x_B(t), x_A(t)x_B(t), 1/x_A(t), 1/x_B(t)]'$, where the inverse terms are included for a better approximation of the Michaelis-Menten kinetics, and the mixed term is included for a better modeling of heterodimer effects. We refer to this extension of the HBR model as the non-linear HBR (HBR-nl) model.

2.6 Non-homogeneous hierarchical Bayesian models

In our applications, described in Section 3, the underlying regulatory relationships are non-linear and vary in dependence on the external light condition. We therefore follow Grzegorczyk and Husmeier (2012) and

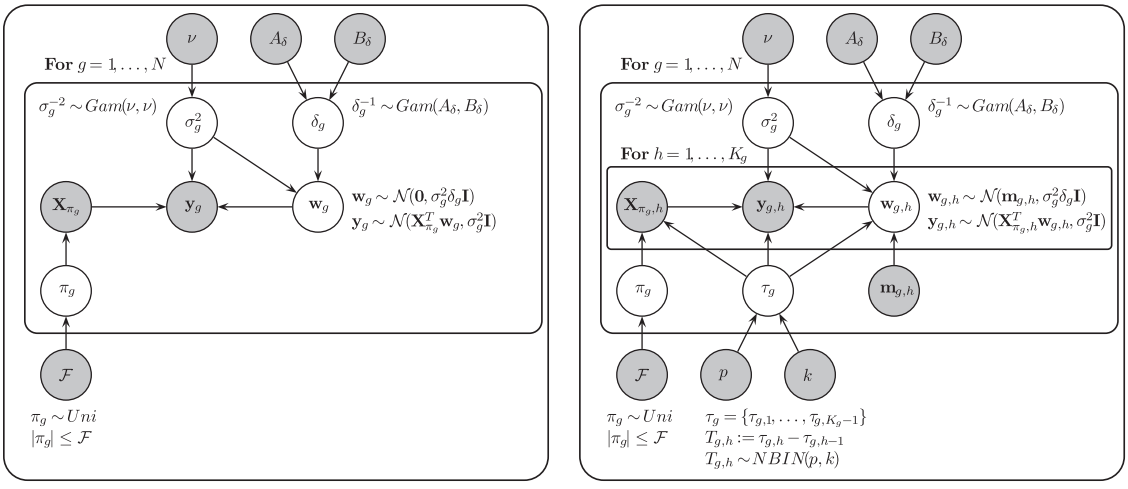

Figure 1 Representation of the hierarchical Bayesian regression models as graphical models.
In both panels the gray circles refer to fixed hyperparameters, while the white circles refer to flexible (hyper-)parameters, which are inferred from the posterior distribution with MCMC. Left panel: The homogeneous Bayesian regression model. The outer plate includes the complete model, and the centre plate refers to the target variables, g=1,…,N; see Section 2.5 for a detailed model description. Right panel: The uncoupled variant of the non-homogeneous Bayesian regression model. Variable-specific change-point sets, $\boldsymbol{\tau}_g$, divide the data into variable-specific segments. The additional inner plate refers to the variable-specific segments, h=1,…,K_g; see Section 2.6 for more details. In the coupled variant of the non-homogeneous Bayesian regression model (not shown in this figure) the regression parameter vectors, $\mathbf{w}_{g,h}$ (h=1,…,K_g), are sequentially coupled via equations (21–22).

combine the Bayesian regression model from Section 2.5 with a multiple change-point process. The change-point process imposes a set of $\mathcal{K}_g - 1$ change-points, $\{\tau_{g,h}\}_{1\leq h\leq(\mathcal{K}_g-1)}$ with $\tau_{g,h}<\tau_{g,h+1}$, to divide the temporal observations of a variable into $\mathcal{K}_g$ disjunct segments. With the two pseudo-change-points $\tau_{g,0}:=1$ and $\tau_{g,\mathcal{K}_g}:=T$ each segment $h\in\{1,\ldots,\mathcal{K}_g\}$ is defined by two demarcating change-points, $\tau_{g,h}$ and $\tau_{g,h+1}$. The vector of the target variable realizations, $\mathbf{y}_g=(y_{g,1},\ldots,y_{g,T})^\top$, is thus divided into $\mathcal{K}_g$ subvectors, $\{y_{g,h}\}_{h=1,\ldots,\mathcal{K}_g}$, where each subvector corresponds to a temporal segment: $\mathbf{y}_{g,h}=(y_{g,(\tau_{g,h}+1)},\ldots,y_{g,\tau_{g,h+1}})^\top$. In Grzegorczyk and Husmeier (2012) the distances between two successive change-points, $T_{g,h}=\tau_{g,h+1}-\tau_{g,h}$, are assumed to have a negative binomial distribution, symbolically $T_{g,h}\sim NBIN(p,k)$; see Grzegorczyk and Husmeier (2012) for the technical details and see Section 2.6.2 for our slightly different implementation.

We keep the covariate set, $\boldsymbol{\pi}_g$, fixed among the $\mathcal{K}_g$ segments, and we apply the linear Gaussian regression model, defined in equation (11), to each segment h:

$$\mathbf{y}_{g,h}|(\mathbf{X}_{\boldsymbol{\pi}_g,h},\mathbf{w}_{g,h},\sigma_g^2)\sim\mathcal{N}(\mathbf{X}_{\boldsymbol{\pi}_g,h}^\top\mathbf{w}_{g,h},\sigma_g^2\mathbf{I}) \tag{20}$$

where $\mathbf{x}_{\pi_g,h}$ is the segment-specific (restricted) design matrix, which can be built from the realizations of the covariate set $\boldsymbol{\pi}_g$ in segment h, and $\mathbf{w}_{g,h}$ is the vector of the segment-specific regression parameters for segment h. As in Section 2.5 we impose an inverse gamma prior on σ_g^2, symbolically $\sigma_g^{-2}\sim\text{Gam}(\nu,\nu)$. For the segment-specific regression parameters, $\mathbf{w}_{g,h}$ $(h=1,\ldots,\mathcal{K}_g)$, we assume Gaussian priors:

$$\mathbf{w}_{g,h}|(\mathbf{m}_{g,h-1},\sigma_g,\delta_g,\mathbf{X}_{\boldsymbol{\pi}_g,h})\sim\mathcal{N}(\mathbf{m}_{g,h-1},\delta_g\sigma_g^2\mathbf{I}) \tag{21}$$

with the hyperprior $\delta_g^{-1}\sim\text{Gam}(A_\delta,B_\delta)$. As with Grzegorczyk and Husmeier (2012) we distinguish two variants of the non-homogeneous Bayesian regression model. In the *uncoupled variant* we set $\mathbf{m}_{g,h}=\mathbf{0}$ for all $h\geq 0$. In the *sequentially coupled variant* we allow for information-sharing between the regression parameters of adjacent segments by setting $\mathbf{m}_{g,0}=\mathbf{0}$, and for $h\geq 1$:

$$\mathbf{m}_{g,h}=\boldsymbol{\Sigma}_{g,h}(\delta_g^{-1}\mathbf{w}_{g,(h-1)}+\mathbf{X}_{\boldsymbol{\pi}_g,h}\mathbf{y}_{g,h}). \tag{22}$$

with $\boldsymbol{\Sigma}_{g,h}^{-1}=\delta_g^{-1}\mathbf{I}+\mathbf{X}_{\boldsymbol{\pi}_g,h}\mathbf{X}_{\boldsymbol{\pi}_g,h}^\top$. As in the previous section, posterior inference is carried out with the Metropolis-Hastings within partially collapsed Gibbs sampling scheme from Grzegorczyk and Husmeier (2012). The marginal likelihood in equation (16) has to be replaced by:

$$P(\boldsymbol{\pi}_g|\mathbf{X}_g,\delta_g,\{\tau_{g,h}\}_{1\leq h\leq(\mathcal{K}_g-1)})\propto P(\boldsymbol{\pi}_g)\prod_{h=1}^{\mathcal{K}_g}p(\mathbf{y}_{g,h}|\mathbf{X}_{\boldsymbol{\pi}_g,h},\delta_g) \tag{23}$$

where $p(\mathbf{y}_{g,h}|\mathbf{X}_{\boldsymbol{\pi}_g,h},\delta_g)$ $(h=1,\ldots,\mathcal{K}_g)$ can be computed in closed-form; see Grzegorczyk and Husmeier (2012) for a mathematical derivation. The full conditional distribution of $\mathbf{w}_{g,h}$ is now given by (Grzegorczyk and Husmeier, 2012):

$$\mathbf{w}_{g,h}|(\mathbf{y}_{g,h},\mathbf{X}_{\boldsymbol{\pi}_g,h},\sigma_g^2,\delta_g)\sim\mathcal{N}(\tilde{\mathbf{m}}_{g,h},\sigma_g^2\boldsymbol{\Sigma}_{g,h}) \tag{24}$$

where $|\boldsymbol{\pi}_g|$ is the cardinality of the covariate set $\boldsymbol{\pi}_g$, and $\boldsymbol{\Sigma}_{g,h}$ was defined below equation (22). For the uncoupled variant of the model we have: $\tilde{\mathbf{m}}_{g,h}=\boldsymbol{\Sigma}_{g,h}\mathbf{X}_{\boldsymbol{\pi}_g,h}\mathbf{y}_{g,h}$. For the coupled variant of the model we have: $\tilde{\mathbf{m}}_{g,h}:=\mathbf{m}_{g,h}$, where $\mathbf{m}_{g,h}$ was defined in equation (22). The full conditional distribution of δ_g^{-1}, symbolically $p(\delta_g^{-1}|\sigma_g^2,\{\mathbf{w}_{g,h}\}_{h=1,\ldots,\mathcal{K}_g})$, is a gamma distribution whose closed-form solution can be found in Grzegorczyk and Husmeier (2012). The inverse variance hyperparameters, σ_g^{-2}, can again be sampled with a collapsed Gibbs sampling step (Grzegorczyk and Husmeier, 2012):

$$\sigma_g^{-2}|(\mathbf{y}_g,\mathbf{X}_{\boldsymbol{\pi}_g},\delta_g,\{\tau_{g,h}\}_{1\leq h\leq(\mathcal{K}_g-1)})\sim\text{Gam}\left(\nu+\frac{T}{2},\nu+\frac{\sum_{h=1}^{\mathcal{K}_g}\Delta_{g,h}^2}{2}\right) \tag{25}$$

with $\Delta^2_{g,h} := (\mathbf{y}_{g,h} - \mathbf{X}_{\boldsymbol{\pi}_g,h}\mathbf{m}_{g,h-1})^\top (\mathbf{I} + \delta_g \mathbf{X}^\top_{\boldsymbol{\pi}_g,h}\mathbf{X}_{\boldsymbol{\pi}_g,h})^{-1} (\mathbf{y}_{g,h} - \mathbf{X}_{\boldsymbol{\pi}_g,h}\mathbf{m}_{g,h-1})$, where $\mathbf{m}_{g,h-1}$ can be computed with equation (22) in the coupled variant, and $\mathbf{m}_{g,h-1} = \mathbf{0}$ for all $h \geq 0$ in the uncoupled variant. A compact graphical representation of the relationships among the (hyper-)parameters of the uncoupled variant of the non-homogeneous Bayesian regression model can be found in the right panel of Figure 1.[7]

Combining the linear regression model with a change-point process provides a natural mechanism to allow for temporal (longitudinal) relationships in the data. However, the data in our study are a mixture of short time series from several independent experiments, where the overall temporal factor influencing the system is the light phase. In addition, we aim to draw on the change-point process as a mechanism to approximate the intrinsic non-linearities of the Michaelis-Menten kinetics via a piece-wise linear model. We therefore treat the data as independent interchangeable realizations and regroup them prior to the application of the change-point process, as explained in the following two subsections.

2.6.1 Fixed change-point induced by the external light condition (HBR-light)

Since light may have a substantial effect on the regulatory relationships of the circadian clock, we divide the observations of the target variables into two segments according to a binary light phase indicator: $h=1$ (light) versus $h=2$ (darkness). This reflects the nature of the laboratory experiments, where *A. thaliana* seedlings are grown in an artificial light chamber whose light is switched on or off. It is straightforward to generalize this approach to more than two segments to allow for extended dawn and dusk periods in natural light. Given that the light phase is known, we consider the segmentation as fixed, and we refer to the model as the hierarchical Bayesian regression (HBR) model with two light-induced components (HBR-light). Since we also assume that light has a substantial influence, we do not penalize any differences between the interaction parameters associated with the two light phases and apply the *uncoupled* non-homogeneous Bayesian regression model, shown in the right panel of Figure 1.

2.6.2 Change-points in the amplitude of the target variable (HBR-cps)

To approximate the non-linear dynamics of the Michaelis-Menten kinetics, we sort the realizations $y_{g,1},\ldots,y_{g,T}$ of each target variable, y_g, in increasing order to obtain the order statistics $y_{g,(1)} \leq \ldots \leq y_{g,(T)}$.[8] Applying the non-homogeneous Bayesian regression models to the ordered realizations, $y_{g,(1)},\ldots,y_{g,T}$, then effectively yields a segmentation of the realizations, $y_{g,1},\ldots,y_{g,T}$, with respect to the amplitude of the target variable y_g. To infer the number of change-points and the change-point locations, we again follow Grzegorczyk and Husmeier (2012) and use a point process prior, where the distance between two successive change-points, $T_{g,h} = \tau_{g,h+1} - \tau_{g,h}$, is assumed to have a negative binomial distribution with hyperparameters $p \in [0, 1]$ and $k=1$, symbolically $T_{g,h} \sim NBIN(p, 1)$. We apply both variants of the non-homogeneous Bayesian regression model. The uncoupled variant is shown in the right panel of Figure 1, and we set $\mathbf{m}_{g,h} = \mathbf{0}$ for all $h \geq 0$ in equation (21). In the coupled variant the regression parameter vectors, $\mathbf{w}_{g,h}$ $(h=1,\ldots,\mathcal{K}_g)$, are sequentially coupled via equations (21–22). We refer to these hierarchical Bayesian regression models as the change-point-divided hierarchical Bayesian regression models (HBR-cps).

2.6.3 Marginal interaction posterior probabilities

For the four previously described hierarchical Bayesian regression models (HBR, HBR-nl, HBR-light, and HBR-cps) MCMC simulation techniques are employed to generate samples from the posterior distribu-

7 We note that the coupled variant of the non-homogeneous Bayesian regression model cannot be represented properly as a graphical model, as the regression parameter vectors are *sequentially* coupled among adjacent segments via equations (21–22).

8 For each y_g we apply exactly the same permutation to order the realizations of the explanatory variables (covariates) and thereby ensure that the segment-specific design matrices are built properly.

tions. Keeping only the sampled regulator sets, $\boldsymbol{\pi}_g^{(1)},\ldots,\boldsymbol{\pi}_g^{(H)}$, corresponds to a marginalization over all other sampled parameters. An estimator of the marginal posterior probability of a regulatory interaction between the regulator x_i^g and the target variable y_g, symbolically $x_i^g \to y_g$, is then given by the fraction of regulator sets that contain x_i^g:

$$P(x_i^g \to y_g) = \frac{1}{H}\sum_{h=1}^{H} I(x_i^g \in \boldsymbol{\pi}_g^{(h)}) \tag{26}$$

where $I(x_i^g \in \boldsymbol{\pi}_{g,h})$ is an indicator function, which is 1 if x_i^g is in the set of regulators $\boldsymbol{\pi}_{g,h}$, and zero otherwise. For the hierarchical Bayesian regression models we use the marginal interaction posterior probabilities to score the interactions with respect to their strengths.

2.7 Automatic relevance determination (ARD-SBR)

The method of automatic relevance determination (ARD) in the context of sparse Bayesian regression (SBR) was proposed by Tipping (2001), and we refer to this method as ARD-SBR. ARD-SBR was first applied to learning gene regulation networks by Rogers and Girolami (2005). It is related to the Bayesian regression method discussed in Section 2.5, with the following modification of the prior on the regression parameters $\mathbf{w}_g$: equation (12) is replaced by

$$p(\mathbf{w}_g \mid \boldsymbol{\alpha}_g) = \mathcal{N}(\mathbf{0},\, \mathrm{diag}[\boldsymbol{\alpha}_g]^{-1}) \tag{27}$$

where $\boldsymbol{\alpha}_g$ is a vector of interaction hyperparameters of the same dimension as $\mathbf{w}_g$, and diag[$\boldsymbol{\alpha}_g$] is a diagonal matrix with $\boldsymbol{\alpha}_g$ in the diagonal. The marginal likelihood, equation (14), now becomes

$$p(\mathbf{y}_g \mid \mathbf{X}_g, \sigma_g^2, \boldsymbol{\alpha}_g) = \int p(\mathbf{y}_g \mid \mathbf{X}_g, \sigma_g^2, \mathbf{w}_g)\, p(\mathbf{w}_g \mid \boldsymbol{\alpha}_g)\, d\mathbf{w}_g = \mathcal{N}(\mathbf{y}_g \mid \mathbf{0},\, \sigma_g^{-2}\mathbf{I} + \mathbf{X}_g^{\top}\mathrm{diag}[\boldsymbol{\alpha}_g]^{-1}\mathbf{X}_g) \tag{28}$$

and is optimized with respect to the hyperparameters $\boldsymbol{\alpha}_g$ in a maximum likelihood type-II manner.[9] Note that as opposed to equation (14), equation (28) depends on the full design matrix $\mathbf{X}_g$, not the design matrix restricted to a subset of regulators $\boldsymbol{\pi}_g$, $\mathbf{X}_{\boldsymbol{\pi}_g}$, and the discrete search in structure space, $\boldsymbol{\pi}_g$, is replaced by a continuous search in hyperparameter space, $\boldsymbol{\alpha}_g$, which is much faster. Hyperparameters $\alpha_{g,i}$ associated with irrelevant regulators x_i^g will be driven to $\alpha_{g,i} \to \infty$, as explained in Section 13.7 of Murphy (2012). The consequence is that the associated regression parameters will be driven to zero, $w_{g,i} \to 0$, and irrelevant regulators x_i^g will effectively be pruned; hence the name "automatic relevance determination" (ARD). For fixed values of the hyperparameters, the posterior of the regression parameters $\mathbf{w}_g$ can be obtained, and the method was therefore originally called "sparse Bayesian regression" (SBR). However, as opposed to the proper Bayesian method discussed in Section 2.5, SBR-ARD is only "Bayesian" about the values of the regression parameters $\mathbf{w}_g$ and does not reflect any uncertainty about $\boldsymbol{\alpha}_g$, which is typically of more interest. Hence, in comparison with Section 2.5, SBR-ARD gains computational speed at the expense of less thorough, approximate inference. How does SBR-ARD compare with the sparse regression methods of Section 2.3? As shown in section 5 of Tipping (2001), the interaction parameters $\boldsymbol{\alpha}_g$ can in principle be integrated out analytically (although this is not advisable for computational reasons). The resulting prior distribution of the regression parameters is $p(w_{g,i}) \propto \frac{1}{|w_{g,i}|}$, where $w_{g,i}$ is the i-th element of the regression parameter vector $\mathbf{w}_g$. The latter prior has more probability mass for $w_{g,i} \to 0$ than the Lasso prior, $p(w_{g,i}) \propto \exp(-|w_{g,i}|)$. Hence, SBR-ARD will lead to sparser network structures than Lasso. As for Lasso, we use the absolute values of the elements of the estimated regression parameter vectors, $\hat{\mathbf{w}}_g$, to score the regulatory effects on the target variable y_g (g=1,…,G) with respect to their strengths.

9 In our study we follow Rogers and Girolami (2005) and use a slightly modified version of the fast marginal likelihood algorithm from Tipping et al. (2003) for optimization.

2.8 Bayesian spline autoregression (BSA)

The Bayesian spline autoregression method (BSA) proposed by Morrissey et al. (2011) is related to the hierarchical Bayesian regression method of Section 2.5 with the essential difference that in the restricted design matrix $\mathbf{X}_{\pi_g}$ the original covariates are augmented with m B-spline basis functions of degree l defined over a set of k evenly spaced knots, where (m, l, k) are user-defined parameters. Consequently, the strength of the interaction between a regulator x_i^g and the target variable y_g, which was modeled with a scalar in the method of Section 2.5, now becomes a vector. That is, each individual element $w_{g,i}$ of the regression parameter vector $\mathbf{w}_g=(w_{g,0}, w_{g,1},\ldots,w_{g,\mathrm{Gg}},)^\top$, where i=0 corresponds to the intercept, is substituted for a vector $\mathbf{w}_{g,i}$, spanning the entire range of B-spline basis functions. To deal with the increased dimension of the resulting total parameter vector $\mathbf{w}_g:=(\mathbf{w}_{g,0}^\top, w_{g,1}^\top,\ldots,w_{g,G_g}^\top)^\top$ and encourage network sparsity, a slab-and-stick-like Bayesian variable selection scheme, first proposed by Smith and Kohn (1996), is used. Define $\mathbf{w}_{g,i}=\gamma_{g,i}\mathbf{u}_{g,i}$, where $\gamma_{g,i}\in\{0, 1\}$ is a binary variable to indicate whether the interaction $x_i^g\to y_g$ is on ($\gamma_{g,i}$=1) or off ($\gamma_{g,i}$=0). The indicator variables $\gamma_{g,i}$ are given a Beta-Bernoulli prior, meaning a Bernoulli prior on $\gamma_{g,i}$ with hyperparameters from a Beta distribution. The higher-level hyperparameters of the Beta distribution have a Jeffreys prior. The parameter vectors $\mathbf{u}_{g,i}$ are given a Gaussian prior to shrink them towards the origin:

$$p(\mathbf{u}_{g,i}|\tau_{g,i})=\mathcal{N}(\mathbf{u}_{g,i}|\mathbf{0}, \tau_{g,i}\mathbf{K})$$

where the structure of the covariance matrix $\mathbf{K}$ is constructed from the second-order differences between adjacent coefficients, and $\tau_{g,i}$ is a smoothness hyperparameter that defines the trade-off between fitting an interpolating spline ($\tau_{g,i}\to 0$) and a straight line ($\tau_{g,i}\to\infty$). Several priors for $\tau_{g,i}$ were tested by Morrissey et al. (2011), with the best performance achieved with an inverted Pareto distribution. Like for the hierarchical Bayesian regression method of Section 2.5, there is no closed-form expression for the posterior distribution, and MCMC sampling based on a Metropolis-within-Gibbs scheme is used: the technical details can be found in Morrissey et al. (2011). The resulting MCMC samples $\gamma_{g,i}^{(1)},\ldots,\gamma_{g,i}^{(H)}$ (g=1,…,G and i=1,…,G_g) are used to estimate the marginal posterior probability of the regulatory interactions $x_i^g\to y_g$:

$$P(x_i^g\to y_g)=\frac{1}{H}\sum_{h=1}^{H}\gamma_{g,i}^{(h)} \tag{29}$$

For the Bayesian spline autoregression method we use these marginal interaction posterior probabilities to score the regulatory interactions with respect to their strengths. The method was originally designed for time series data of the form of equation (3). However, as already discussed at the beginning of Section 2, the underlying approximation equation (2) might be sub-optimal. For a fair comparison with the other methods, we have therefore applied it to target variables $y_{g,t}$ of the form of equation (1).

2.9 State-space models (SSM)

The state-space model (SSM) proposed by Beal et al. (2005) is a Kalman filter with additional Markovian dependence among the observation vectors, and additional dependence of the latent vectors on the observation vectors from the previous time point; see equations (6–7) in Beal et al. (2005). The parameters are estimated with variational Bayesian inference; since all distributions are multivariate Gaussian, this gives closed-form update equations that are carried out iteratively with a modified version of the expectation maximization algorithm. From these parameters, interaction strengths among the genes can be derived; see equation (8) in Beal et al. (2005) for an explicit expression. The interactions contain two separate contributions: direct interactions, describing how gene expression values at the previous time point influence the current expression values, and indirect interactions, modeling gene interactions mediated via the unobserved latent factors. The dimension of the latent vector is unknown and needs to be set using cross-validation or an estimate of the lower bound on the marginal likelihood. The intrinsic Markovian nature of the SSM from Beal et al. (2005) is consistent

with equation (3), but not with equation (1). However, a modification to our data format is straightforward by reverting to an alternative form of the SSM, proposed in Beal (2003, chapter 5), and shown in Figure 2. In fact, the model in Beal et al. (2005) is equivalent to the one in Beal (2003), with the external inputs replaced by the previous observations. The mathematical form of the model is as follows:

$$\mathbf{h}_{t+1} = \mathbf{A}\mathbf{h}_t + \mathbf{B}\mathbf{x}_{\cdot,t} + \boldsymbol{\epsilon}_t$$
$$\mathbf{y}_{\cdot,t} = \mathbf{C}\mathbf{h}_t + \mathbf{D}\mathbf{x}_{\cdot,t} + \boldsymbol{\xi}_t$$

The symbols have the following meaning: $\mathbf{y}_{\cdot,t}$ is the vector of all response variables (i.e., mRNA concentration derivatives) at time t. $\mathbf{x}_{\cdot,t}$ is the vector of all potential regulators at time t; these are either mRNA or protein concentrations. $\mathbf{h}_t$ denotes the vector of unknown latent factors at time t. ϵ_t and $\boldsymbol{\xi}_t$ are vectors of iid white Gaussian noise. The parameters of the model are the transition matrices **A**, **B**,**C** and **D**. These parameters are given prior distributions, which depend on further hyperparameters. For the full hierarchical Bayesian model representation, see Beal (2003). As described in Beal (2003), the posterior expectation of the interaction matrix **CB**+**D** can be employed to assess the strengths of the individual network interactions; see Section 4.6.6 for details.

2.10 Gaussian processes (GP)

Gaussian processes provide a popular method in nonparametric Bayesian statistics for defining a prior distribution directly in the function space rather than the parameter space. By definition, a Gaussian process is a collection of random variables, of which any finite subset has a joint Gaussian distribution. For a gene g the process can be fully represented by a mean function $m_g(\cdot)$ and a covariance function $k_g(\cdot,\cdot)$:

$$f_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}) \sim \mathcal{GP}(m_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}), k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})) \tag{30}$$

where $\mathbf{x}_{\boldsymbol{\pi}_g,t}$ and $\mathbf{x}_{\boldsymbol{\pi}_g,t'}$ are vectors of explanatory variables for target gene g; these are the gene expression values of the set of regulators $\boldsymbol{\pi}_g$, and $\mathbf{x}_{\boldsymbol{\pi}_g,t}$, $\mathbf{x}_{\boldsymbol{\pi}_g,t'}$ are the corresponding subsets of $\mathbf{X}_{\boldsymbol{\pi}_g}$; see Table 1 for an overview of the notation. The mean function $m_g(\cdot)$ is usually set to zero, which presents prior ignorance about the trend (i.e., we are equally unsure that a trend is up or down). An important feature of Gaussian processes is that, due to the Gaussianity assumption, marginalization integrals have closed form solutions. In particular, we get for the marginal likelihood, under the assumption of independent and identically distributed additive Gaussian noise with variance σ_g^2 (Rasmussen and Williams, 2006):

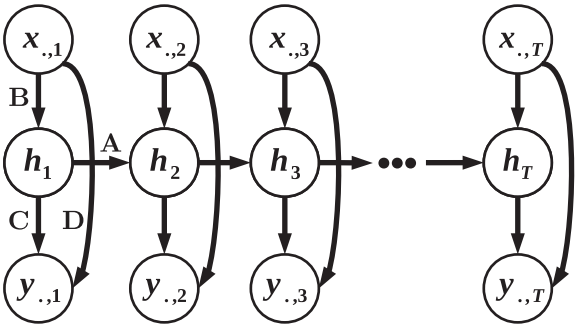

Figure 2 Graphical model representation of the state-space model (SSM).
The figure is adapted from figure 5.2 of Beal (2003). $\mathbf{y}_{\cdot,t}$ represents the vector of all response variables (i.e., mRNA concentration derivatives) at time t. $\mathbf{x}_{\cdot,t}$ represents the vector of all potential regulators at time t; depending on the problem, these are either mRNA or protein concentrations. h_t denotes the vector of unknown latent factors at time t. The arrows indicate probabilistic dependence relations. The parameters of the model are the four transition matrices shown in capital letters **A**, **B**, **C**, and **D**. These parameters are given prior distributions, which depend on further hyperparameters. For the full hierarchical Bayesian model representation, see Beal (2003).

$$p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\theta}_g)=\frac{1}{\sqrt{(2\pi)^T|\mathbf{K}_g+\sigma_g^2\mathbf{I}|}}\exp\left(-\frac{1}{2}\mathbf{y}_g^\top(\mathbf{K}_g+\sigma_g^2\mathbf{I})^{-1}\mathbf{y}_g\right) \tag{31}$$

where $\mathbf{y}_g=(y_{g,1},\ldots,y_{g,T})^\top$ is a vector of target values for gene g, and $\mathbf{K}_g$ is a T-by-T covariance matrix, with elements $K_{g,t,t'}=k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})$. The arguments of the kernel function $k_g(\cdot,\cdot)$ are the vectors of gene expression values associated with the putative regulators of gene g, $\boldsymbol{\pi}_g$, taken at the time points t and t'; these vectors are extracted from the (restricted) design matrix $\mathbf{X}_{\boldsymbol{\pi}_g}$. The kernel function depends on certain hyperparameters $\boldsymbol{\theta}_g$. For the widely applied squared exponential kernel

$$k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})=a_g\exp\left(-\frac{(\mathbf{x}_{\boldsymbol{\pi}_g,t}-\mathbf{x}_{\boldsymbol{\pi}_g,t'})^2}{2l_g^2}\right) \tag{32}$$

these are the length scale l_g and amplitude a_g:$\boldsymbol{\theta}_g=(l_g, a_g)$. In our work we follow Äijö and Lähdesmäki (2009) and choose a Matérn class kernel

$$k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})=a_g\left(1+\sqrt{\frac{3}{l_g^2}}(x_{\boldsymbol{\pi}_g,t}-x_{\boldsymbol{\pi}_g,t'})^\top(x_{\boldsymbol{\pi}_g,t}-x_{\boldsymbol{\pi}_g,t'})\right)\exp\left(-\sqrt{\frac{3}{l_g^2}}(x_{\boldsymbol{\pi}_g,t}-x_{\boldsymbol{\pi}_g,t'})\ (x_{\boldsymbol{\pi}_g,t}-x_{\boldsymbol{\pi}_g,t'})\right) \tag{33}$$

which provides a better compromise between smoothness and roughness. Like for the squared exponential kernel, the hyperparameters $\boldsymbol{\theta}_g$ consist of a length scale and an amplitude parameter: $\boldsymbol{\theta}_g=(l_g, a_g)$. In order to apply Gaussian processes to the inference of gene regulatory networks, we follow the approach described by Äijö and Lähdesmäki (2009). The starting point is the mathematical formulation of transcriptional regulation of equation (1), whose right-hand side can be reformulated as follows:

$$\tilde{f}_g(\mathbf{x}_{\boldsymbol{\pi}_g,t})=f_g(\mathbf{x}_{\boldsymbol{\pi}_g,t})+\mathbf{h}_g^\top\boldsymbol{\beta}_g \tag{34}$$

where $\boldsymbol{\beta}_g=(\alpha_g, \lambda_g)$ and $\mathbf{h}_g=(1, -x_{g,t})$. The approach taken by Äijö and Lähdesmäki (2009) is to impose a normal distribution with mean vector $\mathbf{b}$ and covariance matrix $\mathbf{B}=\sigma_b^2\mathbf{I}$ on $\boldsymbol{\beta}_g$:

$$\boldsymbol{\beta}_g\sim N(\mathbf{b}, \mathbf{B})=N(\mathbf{b}, \sigma_b^2\mathbf{I}) \tag{35}$$

It can then be shown (Rasmussen and Williams, 2006) that a Gaussian process assumption for f_g

$$f_g(\mathbf{x}_{\boldsymbol{\pi}_g,t})\sim\mathcal{GP}(0, k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})) \tag{36}$$

implies a Gaussian process for $\tilde{f}_g$ of the following form:

$$\tilde{f}_g(\mathbf{x}_{\boldsymbol{\pi}_g,t})\sim\mathcal{GP}(\mathbf{h}_g^\top\mathbf{b}, k_g(\mathbf{x}_{\boldsymbol{\pi}_g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t'})+\mathbf{h}_g^\top\mathbf{B}\mathbf{h}_g) \tag{37}$$

This gives, in modification of equation (31), a closed from expression for the marginal likelihood

$$p(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) \tag{38}$$

for which the explicit expression can be obtained from Äijö and Lähdesmäki (2009). Note that the target values $\mathbf{y}_g$ are time derivatives, which Äijö and Lähdesmäki (2009) approximate by difference quotients. The hyperparameters $\boldsymbol{\theta}_g=(a_g, l_g)$ and the noise variance σ_g^2 are optimized so as to maximize the marginal likelihood in equation (38). This can be achieved with the Polack-Ribiere conjugate gradient method, as described by Rasmussen and Williams (2006). To avoid negative values of $\boldsymbol{\beta}_g$, which are biologically implausible, Äijö and Lähdesmäki (2009) suggested setting the hyperparameters $\mathbf{b}$ and σ_b^2 to fixed values such that plausible

values of $\boldsymbol{\beta}_g$ have high probability. To accomplish structure learning for a target variable y_g, the posterior probability for a selected set of regulators, $\boldsymbol{\pi}_g$, can be obtained from Bayes' theorem:

$$P(\boldsymbol{\pi}_g \mid \mathbf{y}_g, \mathbf{X}_g, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) = \frac{p(\mathbf{y}_g \mid \mathbf{X}_{\boldsymbol{\pi}_g}, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) P(\boldsymbol{\pi}_g)}{\sum_{g'} p(\mathbf{y}_{g'} \mid \mathbf{X}_{\boldsymbol{\pi}'_g}, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) P(\boldsymbol{\pi}_{g'})} \tag{39}$$

where $P(\boldsymbol{\pi}_g)$ is the prior probability distribution on the set of potential regulators, for which Äijö and Lähdesmäki (2009) chose a uniform distribution. The posterior probability of a particular gene interaction between the i-th regulator x_i^g and the target y_g is then given by marginalization:

$$P(x_i^g \to y_g \mid \mathbf{y}_g, \mathbf{X}_g, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) = \sum_{\boldsymbol{\pi}_g} I(x_i^g \in \boldsymbol{\pi}_g) P(\boldsymbol{\pi}_g \mid \mathbf{y}_g, \mathbf{X}_{\pi_g}, \boldsymbol{\theta}_g, \sigma_g^2, \mathbf{b}, \sigma_b^2) \tag{40}$$

where $I(x_i^g \in \boldsymbol{\pi}_g)$ is the indicator function, which is 1 if x_i^g is in the set of regulators $\boldsymbol{\pi}_g$, and zero otherwise. For larger networks, where a complete enumeration of all potential sets of regulators is computationally prohibitive, the common approach is to impose a fan-in restriction, e.g., of 3, i.e., $P(\boldsymbol{\pi}_g)=0$ if $|\boldsymbol{\pi}_g|>3$. The posterior distribution of equation (40) can be used to score the regulatory interactions with respect to their strengths. The Matlab software *GP4GRN* from Äijö and Lähdesmäki (2009) implements the described framework and was used in our study.

2.11 Mutual information methods (ARACNE)

Consider three variables x_1, x_2 and x_3. The mutual information (MI) between x_1 and x_2 is then given by

$$I(x_1, x_2) = \int p(x_1, x_2) \log\left[\frac{p(x_1, x_2)}{p(x_1)\,p(x_2)}\right] dx_1 dx_2 \geq 0 \tag{41}$$

$I(x_1, x_2)$ is zero if the expression profiles of x_1 and x_2 are stochastically independent: $p(x_1, x_2)=p(x_1)p(x_2)$. The mutual information measures the degree of stochastic dependence between x_1 and x_1, which in earlier work by Butte and Kohane (2000) was used to provide a ranking of all potential gene interactions. A permutation test can then be used to set a threshold for discarding low-ranked interactions at a specified significance level. A shortcoming of this approach is the fact that direct interactions are not distinguished from indirect ones. Consider, for instance, a chain reaction

$$x_1 \to x_2 \to x_3$$

where gene x_3 is indirectly regulated by gene x_1 via the intermediary x_2, or the joint regulation of genes x_1 and x_3 by gene x_2:

$$x_1 \leftarrow x_2 \to x_3$$

In both scenarios the variables x_1 and x_3 are stochastically dependent, and $I(x_1, x_3)$ may be large despite the fact that there is no actual interaction between x_1 and x_3. To filter out such spurious interactions, a pruning mechanism was proposed by Margolin et al. (2006), which is based on the data processing inequality: for the above interaction scenarios,

$$I(x_1, x_3) \leq \min\{I(x_1, x_2), I(x_2, x_3)\}$$

The proposed algorithm, called ARACNE, visits each gene triplet in turn and removes the interaction with the smallest mutual information score. Each triplet is analyzed irrespectively of whether its interactions have been marked for removal by prior pruning applications to different triplets, making the algorithm invariant with respect to a reordering of the genes. A theoretical analysis of the types of networks that can be reconstructed

with this algorithm can be found in Margolin et al. (2006). The practical problem is related to the fact that equation (41) cannot be computed exactly from a finite sample size but either requires a discretization of the data (information loss), or the approximation of the probability densities $p(\cdot)$ by a kernel density estimator; see Murphy (2012, chapter 14) for details. While the density itself depends critically on the bandwidth of this estimator, the ranking of mutual information scores has been found to be quite robust with respect to a variation of the bandwidth parameter; see Figure 1 in Margolin et al. (2006). To apply ARACNE to gene expression time series, a time delayed version has been proposed by Zoppoli et al. (2010), which can deal with dynamic processes in the form of equation (3). However, as already discussed at the beginning of Section 2, the underlying approximation equation (2) might be sub-optimal, and we therefore apply ARACNE directly to equation (1). That is, we apply ARACNE to each target variable y_g and its potential regulators $x_1^g,\ldots,x_{G_g}^g$ separately, to obtain mutual interaction scores $I^A(y_g, x_j^g)$ (j=1,…,G_g), where $I^A(y_g, x_j^g)=I(y_g, x_j^g)$ or $I^A(y_g, x_j^g)=0$ if the interaction has been pruned by the ARACNE algorithm. The ARACNE mutual interaction scores can then be interpreted in a bipartite manner, i.e., $I^A(y_g, x_j^g)$ is the strength of the regulatory interaction $x_j^g \to y_g$ (g=1,…,G and j=1,…,G_g).

2.12 Mixture Bayesian network models (MBN)

A flexible Gaussian mixture model approach for inferring non-linear network interactions has been proposed by Ko et al. (2007, 2009), which they call the "Mixture Bayesian network model."[10] The key idea is to model each target gene g conditional on its regulators in $\boldsymbol{\pi}_g$ with a conditional Gaussian mixture model. Given the vector of the variables in a regulator set $\boldsymbol{\pi}_g$ at time t, symbolically $\mathbf{x}_{\boldsymbol{\pi}_g,t}$, we consider a Gaussian mixture model with $\mathcal{K}_g$ mixture components and the mixture weights $\alpha_{g,1},\ldots,\alpha_{g,\mathcal{K}_g}$ for the joint distribution of the target gene $y_{g,t}$ and its regulators $\mathbf{x}_{\boldsymbol{\pi}_g,t}$. Recalling the definition $\mathbf{z}_{\boldsymbol{\pi}_g,t}:=(y_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t}^\top)^\top$ from Table 1 we obtain:

$$p(\mathbf{z}_{\boldsymbol{\pi}_g,t})=\sum_{h=1}^{\mathcal{K}_g}\alpha_{g,h}f_{g,h}(\mathbf{z}_{\boldsymbol{\pi}_g,t}) \tag{42}$$

where each component-specific function $f_{g,h}(\cdot)$ is the density function of a ($|\boldsymbol{\pi}_g|$+1)-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_{g,h}$ and covariance matrix $\boldsymbol{\Sigma}_{g,h}$, and $\sum_{h=1}^{\mathcal{K}_g}\alpha_{g,h}=1$. The marginal distribution of the vector $\mathbf{x}_{\boldsymbol{\pi}_g,t}$ is then also a Gaussian mixture:

$$p(\mathbf{x}_{\boldsymbol{\pi}_g,t})=\sum_{h=1}^{\mathcal{K}_g}\alpha_{g,h}f^*_{g,h}(\mathbf{x}_{\boldsymbol{\pi}_g,t}) \tag{43}$$

where the $|\boldsymbol{\pi}_g|$-dimensional Gaussian density functions $f^*_{g,h}(\cdot)$ have mean vectors $\boldsymbol{\mu}^*_{g,h}$ and covariance matrices $\boldsymbol{\Sigma}^*_{g,h}$ which are sub-vectors of $\boldsymbol{\mu}_{g,h}$ and sub-matrices of $\boldsymbol{\Sigma}_{g,h}$, respectively.[11] Considering $\mathbf{z}_{\pi_g,t}$ (t=1,…,T) as an i.i.d. sample and taking into account that the conditional distribution $p(y_{g,t}|\mathbf{x}_{\boldsymbol{\pi}_g,t})$ is the ratio of the joint distribution in equation (42) and the marginal distribution in equation (43), the likelihood of the conditional Gaussian mixture model is given by:

$$LL(\mathbf{y}_g|\mathbf{X}^*_{\boldsymbol{\pi}_g}, \boldsymbol{\theta}(\boldsymbol{\pi}_g, \mathcal{K}_g))=\frac{\prod_{t=1}^{T}\sum_{h=1}^{\mathcal{K}_g}\alpha_{g,h}f_{g,h}(\mathbf{z}_{\boldsymbol{\pi}_g,t})}{\prod_{t=1}^{T}\sum_{h=1}^{\mathcal{K}_g}\alpha_{g,h}f^*_{g,h}(x_{\boldsymbol{\pi}_g,t})} \tag{44}$$

where $\boldsymbol{\theta}(\boldsymbol{\pi}_g, \mathcal{K}_g)$ denotes the set of mixture parameters, namely the mixture weights as well as the mean vectors and covariance matrices of the component-specific Gaussian distributions, $\mathbf{X}^*_{\boldsymbol{\pi}_g}$ is the matrix of the observations of the regulators in $\boldsymbol{\pi}_g$, and $\mathbf{y}_g=(y_{g,1},\ldots,y_{g,T})^\top$ is the vector of the target variable observations.

10 We use the authors' terminology, although the model is not a proper Bayesian network.

11 More precisely, $\boldsymbol{\mu}^*_{g,h}$ is obtained by deleting the element corresponding to the target variable $y_{g,t}$ in $\boldsymbol{\mu}_{g,h}$, and $\boldsymbol{\Sigma}^*_{g,h}$ is obtained by deleting the row and the column corresponding to $y_{g,t}$ in $\boldsymbol{\Sigma}_{g,h}$.

Given a fixed set of regulators, $\boldsymbol{\pi}_g$, and a fixed number of mixture components, $\mathcal{K}_g$, the maximum likelihood (ML) estimates for the mixture parameters $\boldsymbol{\theta}(\boldsymbol{\pi}_g, \mathcal{K}_g)$ can be obtained with the Expectation-Maximization (EM) algorithm, as described in detail by Ko et al. (2009). Keeping $\boldsymbol{\pi}_g$ fixed, ML estimates, $\hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_g)$, can be computed for different numbers of mixture components $\mathcal{K}_g$. Having estimates $\hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_g)$ for $\mathcal{K}_g=1,\ldots,\mathcal{K}_{MAX}$, where $\mathcal{K}_{MAX}=10$ is an imposed upper bound on the number of mixture components, the Bayesian Information Criterion (BIC) is employed to determine the *best* number of mixture components given $\boldsymbol{\pi}_g$:

$$\mathcal{K}_{g|\boldsymbol{\pi}_g}^{BIC} = \underset{\mathcal{K}_g}{\operatorname{argmin}}\{-2\log(LL(\mathbf{y}_g|\mathbf{X}_{\boldsymbol{\pi}_g}^{\star}, \hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_g)))+\log(T)|\hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_g)|\} \tag{45}$$

where T is the number of observations, $|\hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_g)|$ is the number of the ML-estimated mixture parameters and the likelihood $LL(\cdot|\cdot)$ has been defined in equation (44). With equation (45) the best number of mixture components $\mathcal{K}_{g|\boldsymbol{\pi}_g}^{BIC}$ can be determined for each potential regulator set $\boldsymbol{\pi}_g$. In our implementation we systematically compute $\mathcal{K}_{g|\boldsymbol{\pi}_g}^{BIC}$ for each set $\boldsymbol{\pi}_g$ with a cardinality $|\boldsymbol{\pi}_g|\leq 3$. Finally, the best set of regulators $\boldsymbol{\pi}_g^{BIC}$ for target variable y_g minimizes the BIC criterium, and is thus given by:

$$\boldsymbol{\pi}_g^{BIC} = \underset{\boldsymbol{\pi}_g}{\operatorname{argmin}}\{-2\log(LL(\mathbf{y}_g|\mathbf{x}_{\boldsymbol{\pi}_g}, (\boldsymbol{\pi}_g, \mathcal{K}_{g|\boldsymbol{\pi}_g}^{BIC})))+\log(T)|\hat{\boldsymbol{\theta}}(\boldsymbol{\pi}_g, \mathcal{K}_{g|\boldsymbol{\pi}_g}^{BIC})|\} \tag{46}$$

We repeat the optimization procedure, outlined above, several times and we average over the obtained results, as described in Section 4.6.9, to score the individual interactions, $x_i^g \to y_g$.

2.13 Gaussian Bayesian networks (BGe)

The BGe scoring metric was introduced by Geiger and Heckerman (1994) and has become a standard modeling framework for static and dynamic Gaussian Bayesian networks.[12] For $t=1,\ldots,T$ the common distribution of the target variable $y_{g,t}$ and its potential regulators $\mathbf{x}_{g,t}$ is assumed to be an i.i.d. sample from a (G_g+1)-dimensional multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$p(\mathbf{z}_{g,t}|\boldsymbol{\mu}, \Sigma)=(2\pi)^{-\left(\frac{G_g+1}{2}\right)}\det(\Sigma)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(z_{g,t}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{z}_{g,t}-\boldsymbol{\mu})\right\} \tag{47}$$

where G_g is the number of potential regulators for the target variable y_g, i.e., the length of the vector $\mathbf{x}_{\boldsymbol{\pi}_g,t}$, and $\mathbf{z}_{g,t}:=(y_{g,t}, \mathbf{x}_{\boldsymbol{\pi}_g,t}^{\top})^{\top}$, as defined in Table 1. Onto the unknown parameters, namely the mean vector $\boldsymbol{\mu}$ and the precision matrix $\mathbf{W}:=\Sigma^{-1}$, a normal-Wishart prior is imposed, symbolically:

$$p(\mathbf{W}|\alpha, \mathbf{T}_0)=c(G_g, \alpha)\det(\mathbf{T}_0)^{\frac{\alpha}{2}}\det(\mathbf{W})^{\frac{\alpha-G_g-1}{2}}\exp\left\{-\frac{1}{2}\operatorname{trace}(\mathbf{T}_0\mathbf{W})\right\} \tag{48}$$

$$p(\mu|\mu_0, (\nu\mathbf{W})^{-1})=(2\pi)^{-\left(\frac{G_g+1}{2}\right)}\det(\nu\mathbf{W})^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\top}\nu W(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\} \tag{49}$$

where

$$c(G_g, \alpha)=\left(2^{\frac{\alpha(G_g+1)}{2}}\pi^{\frac{(G_g+1)G_g}{4}}\prod_{i=1}^{G_g+1}\Gamma\left(\frac{\alpha+1-i}{2}\right)\right)^{-1} \tag{50}$$

and the hyperparameters α, $\mathbf{T}_0$, ν and $\boldsymbol{\mu}_0$ of the normal-Wishart distribution are chosen fixed. Geiger and Heckerman (1994) show that the marginal likelihood:

12 Note that the abbreviation "BGe" was introduced by Geiger and Heckerman (1994) and stands for *Bayesian metric for Gaussian networks having score equivalence*; see Geiger and Heckerman (1994) for more details.

$$p(\mathbf{z}_{g,1}, \ldots, \mathbf{z}_{g,T}) = \iint \left(\left\{ \prod_{t=1}^{T} p(\mathbf{z}_{g,t} | \boldsymbol{\mu}, \Sigma) \right\} p(\boldsymbol{\mu} | \boldsymbol{\mu}_0, (\nu \mathbf{W})^{-1}) p(\mathbf{W} | \alpha, \mathbf{T}_0) \right) d\boldsymbol{\mu} d\mathbf{W} \tag{51}$$

can then be computed in closed-form. If it is further assumed that the target variable y_g, conditional on the set of regulators $\boldsymbol{\pi}_g$, becomes statistically independent of all the other potential regulators, symbolically $p(\mathbf{y}_g | \mathbf{X}_g^*, \boldsymbol{\pi}_g) = p(\mathbf{y}_g | X_{\boldsymbol{\pi}_g}^*)$, then the conditional distributions

$$p(\mathbf{y}_g | \mathbf{X}_g^*, \boldsymbol{\pi}_g) = p(\mathbf{y}_g | \mathbf{X}_{\boldsymbol{\pi}_g}^*) = \frac{p(\mathbf{y}_g, \mathbf{X}_{\boldsymbol{\pi}_g}^*)}{p(\mathbf{X}_{\boldsymbol{\pi}_g}^*)} \tag{52}$$

can also be computed in closed-form for each regulator set $\boldsymbol{\pi}_g$, see Geiger and Heckerman (1994) for details. Imposing uniform priors on the regulator sets, $\boldsymbol{\pi}_g$, subject to a maximal cardinality restriction $\mathcal{F}$, the posterior distribution of the regulator set $\boldsymbol{\pi}_g$ with $|\boldsymbol{\pi}_g| \leq \mathcal{F}$ is given by:

$$P(\boldsymbol{\pi}_g | \mathbf{y}_g, \mathbf{X}_g^*) = \frac{p(\mathbf{y}_g | \mathbf{X}_{\boldsymbol{\pi}_g}^*)}{\sum_{\tilde{\boldsymbol{\pi}}_g : |\tilde{\boldsymbol{\pi}}_g| \leq \mathcal{F}} p(\mathbf{y}_g | \mathbf{X}_{\tilde{\boldsymbol{\pi}}_g}^*)} \tag{53}$$

where the sum in the denominator is over all valid regulator sets $\tilde{\boldsymbol{\pi}}_g$ whose cardinality is lower than or equal to the fan-in $\mathcal{F}$. The posterior probability of an interaction between x_i^g and y_g can then be computed by marginalization:

$$P(x_i^g \to y_g | \mathbf{y}_g, \mathbf{X}_g^*) = \sum_{\boldsymbol{\pi}_g} I(x_i^g \in \boldsymbol{\pi}_g) P(\boldsymbol{\pi}_g | \mathbf{y}_g, \mathbf{X}_g^*) \tag{54}$$

where $I(x_i^g \in \boldsymbol{\pi}_g)$ is the indicator function, which is 1 if x_i^g is in the set of regulators $\boldsymbol{\pi}_g$, and zero otherwise. We use the posterior probabilities in equation (54) to score the regulatory interactions with respect to their strengths.

3 Data

This section describes the data used for a critical comparative assessment of the method performance. We use a combination of *real* laboratory data and *realistic* simulated data.

Real data have the advantage that they were obtained from real organisms using real assays. In our case, these are transcriptional concentration time courses from *A. thaliana* seedlings obtained with quantitative reverse transcription polymerase chain reaction (qRT-PCR). The use of real data mimics the actual application a biologist is interested in. A disadvantage, however, is the absence of a ground truth, making it difficult to evaluate the prediction from the different methods.

Realistic data are simulated from a mathematical model of the molecular interactions occurring in the signaling pathways and regulatory networks. Since the data have been synthetically generated, the ground truth is known and can be used for an objective performance evaluation. A disadvantage is that the data generation process might make simplifying assumptions that render the data insufficiently representative of real biological systems studied in the laboratory. The challenge, hence, is to make the data generation process as realistic as possible, and we describe below how we have accomplished this objective.

3.1 Generation of realistic data

Various mathematical models have been developed to describe the molecular interactions and signal transduction processes in the central circadian clock of *A.thaliana* (Locke et al., 2006; Pokhilko et al., 2012, 2013).

They are based on systems of ordinary differential equations (ODEs) that describe the chemical kinetics of transcription initiation, translation, and post-translational modification, using mass action kinetics and/or Michaelis-Menten kinetics. In principle, we could use these mathematical models together with the published values of the kinetic rate parameters to generate synthetic transcription profiles from the circadian regulatory networks published by Locke et al. (2006) and Pokhilko et al. (2012, 2013), then use the latter as a gold standard for our method evaluation.

However, this approach would not generate data that are sufficiently biologically realistic. The solutions of ODEs typically converge to limit cycles with regular oscillations and constant amplitude, which fail to capture the stochastic amplitude variation observed in real qRT-PCR experiments. In addition, the damping of oscillations experimentally observed in constant light conditions is not correctly modeled. The problem of ODEs is that the intrinsic fluctuations of molecular processes in the cell are ignored, thereby not allowing for molecular noise that may have a significant impact on the behaviour of the system (Wilkinson, 2009; Guerriero et al., 2012).

For a more realistic approach, we model the individual molecular processes of transcription, translation, degradation, dimerization etc. as individual discrete events, as shown in Tables 2 and 3. Statistical mechanics arguments then lead to a Markov jump process in continuous time whose instantaneous reaction rates are directly proportional to the number of molecules of each reacting component (Wilkinson, 2009, 2011). Such dynamics can be simulated exactly using standard discrete-event simulation techniques, as illustrated in Table 2. For our study, we followed Guerriero et al. (2012) and adopted the Bio-PEPA framework (Ciocchetta and Hillston, 2009) to simulate gene expression profiles for the core circadian clock of *A. thaliana*, using the Bio-PEPA Eclipse Plug-in.[13] This framework is built on a stochastic process algebra implementation of chemical kinetics, and the stochastic simulations are run with the Gillespie algorithm (Gillespie, 1977).

In order to correctly quantify stochastic fluctuations, concentrations are represented as numbers of molecules per unit volume. This requires the unit volume size Ω to be defined, which scales the molecule amounts and kinetic laws such that a unit concentration in an ODE representation becomes a molecule count close to Ω; see Guerriero et al. (2012) for more details. The size of Ω has a strong influence on the stochasticity of the system. Since larger volumes entail a more pronounced averaging effect, the stochasticity decreases with increasing values of Ω, and the solutions from the equivalent deterministic ODEs are subsumed as a limiting case for $\Omega \to \infty$. Conversely, decreasing values of Ω increase the stochasticity. Guerriero et al. (2012) showed that replacing the continuous deterministic dynamics of ODEs by the discrete stochastic dynamics with an appropriate choice of Ω leads to a more accurate matching of the experimental data, including the damping of oscillations experimentally observed in constant light, better entrainment to light in several light patterns, better entrainment to changes in photo period, and the correct modeling of secondary peaks experimentally observed for certain photo periods.

Table 2 Illustration of elementary molecular reactions with discrete stochastic kinetics.

Elementary molecular reaction	
$X_{DNA} + X_{protein} \xrightarrow{k_1} X_{DNA} + X_{mRNA} + X_{protein}$	Transcription
$X_{mRNA} \xrightarrow{k_2} X_{mRNA} + X_{protein}$	Translation
$X_{mRNA} \xrightarrow{k_3} \varnothing$, $X_{protein} \xrightarrow{k_4} \varnothing$	Degradation
$2X_{protein} \xrightarrow{k_5} X_{dimer}$	Dimerization

The letter "X" represents a single molecule of the type indicated by the subscript, the symbol ∅ indicates the disappearance of a molecule. Arrows indicate reactions, i.e. the transformation of the products on the left to the products on the right. The lower case letters above the arrows denote chemical kinetic parameters. The reactions are modeled mathematically with a Markov jump process. Reactions occur stochastically according to a Poisson process, whose intensity is the sum of the kinetic parameters; here: $\lambda = k_1 + \ldots + k_5$. The propensity of a reaction is proportional to its kinetic parameter, i.e., given that a reaction has occurred, the probability that the nature of this reaction is of type i is k_i/λ.

13 http://www.biopepa.org.

Table 3 Ordinary differential equations (ODEs) and corresponding discrete molecular reaction kinetics for the morning gene prr9.

Chemical Kinetics Described by Ordinary Differential Equations (ODEs)

mRNA Concentration Change

$$\frac{dPRR9_{mRNA}}{dt} = q_3 \cdot light \cdot P_{protein} + n_7 \cdot \frac{g_8^h}{g_8^h + TOC1_{protein}^h} \cdot \frac{LHY_{protein}^i}{LHY_{protein}^i + g_9^i} - m_{12} \cdot PRR9_{mRNA}$$

Protein Concentration Change

$$\frac{dPRR9_{protein}}{dt} = p_8 \cdot PRR9_{mRNA} - (m_{13} \cdot light + m_{22} \cdot dark) \cdot PRR9_{protein}$$

Discrete Stochastic Kinetics of Molecular Reactions

mRNA Count Update

$$PRR9_{mRNA} = PRR9_{transcr} \uparrow + PRR9_{mRNA.degrad} \downarrow$$

$$PRR9_{transcr} = \Omega \cdot \left(\frac{q_3}{\Omega} \cdot light \cdot P_{protein} + \frac{(g_8 \cdot \Omega)^h}{(g_8 \cdot \Omega)^h + TOC1_{protein}^h} \cdot \left(n_4 + n_7 \cdot \frac{LHY_{protein}^i}{LHY_{protein}^i + (g_9 \cdot \Omega)^i} \right) \right)$$

$$PRR9_{mRNA.degrad} = m_{12} \cdot PRR9mRNA$$

Protein Count Update

$$PRR9_{protein} = PRR9_{translate} \uparrow + PRR9_{protein.degrad} \downarrow$$

$$PRR9_{translate} = p_8 \cdot PRR9_{mRNA}$$

$$PRR9_{protein.degrad} = (m_{13} \cdot light + m_{22} \cdot dark) \cdot PRR9_{protein}$$

The symbol "$PRR9_x$" denotes the concentration of a molecular species of the morning gene *PRR9*, specified by the index "x." For instance, $PRR9_{mRNA}$ is the concentration of mRNA transcribed from *PRR9*, $PRR9_{protein}$ is the concentration of *PRR9* protein, etc. The symbol *light* is a binary indicator for the status of light (1=light, 0=darkness), *dark*=1-*light*, lower case letters indicate kinetic parameters, and Ω is a volume parameter. Top panel: ODE description of chemical kinetics, with non-linear Michaelis-Menten kinetics for mRNA concentration change, and linear mass action kinetics for protein concentration change. Bottom panel: The corresponding discrete kinetic reactions, which in the limit $\Omega \to \infty$ converge to the ODE solutions. An upper arrow $\uparrow$ on the right indicates an amount by which the quantity on the left is increased, a down arrow $\downarrow$ on the right indicates an amount by which the quantity on the left is decreased. The reactions occur stochastically, with propensities determined by the reaction rates. Mathematical details can be found in Wilkinson (2011). The complete set of equations for all genes in the central circadian clock of *A. thaliana* is available from Guerriero et al. (2012).

We simulated mRNA and protein concentration profiles over time from the circadian clock regulatory network published in Guerriero et al. (2012) and Pokhilko et al. (2010), shown in Figure 3 (top left, network "wildtype") and Figure 12 (middle left, network "P2010"). This involves genetic regulatory reactions for mRNA transcription, protein translation, and mRNA and protein degradation for nine genes. Table 3 shows the chemical kinetic reactions for a single gene in this network (PRR9), as an illustration. A full list of reactions and their corresponding mathematical descriptions is available from the supplementary material of Guerriero et al. (2012).

An additional advantage of this procedure is that it is straightforward to assess the effect of network structure modification on the performance of the network reconstruction methods. This can easily be effected by inactivating certain reactions in the gold standard network, by setting the respective reaction rates to zero. Figure 3 shows the complete circadian regulatory network in *A. thaliana*, as published by Guerriero et al. (2012) and Pokhilko et al. (2010) ("wildtype"), and several modified sparser structures, which are used throughout our study. The exact setup of the data generation process is described in detail in Section 4.1.

3.2 Real data

In addition to the realistic data simulated from a faithful mathematical description of the molecular interaction processes, as described above, we used real transcription profiles for the key circadian regulatory genes in the model plant *A. thaliana*. The objective is to infer putative gene regulatory networks with the various statistical methods described in Section 2, and then to compare these predictions with network models of

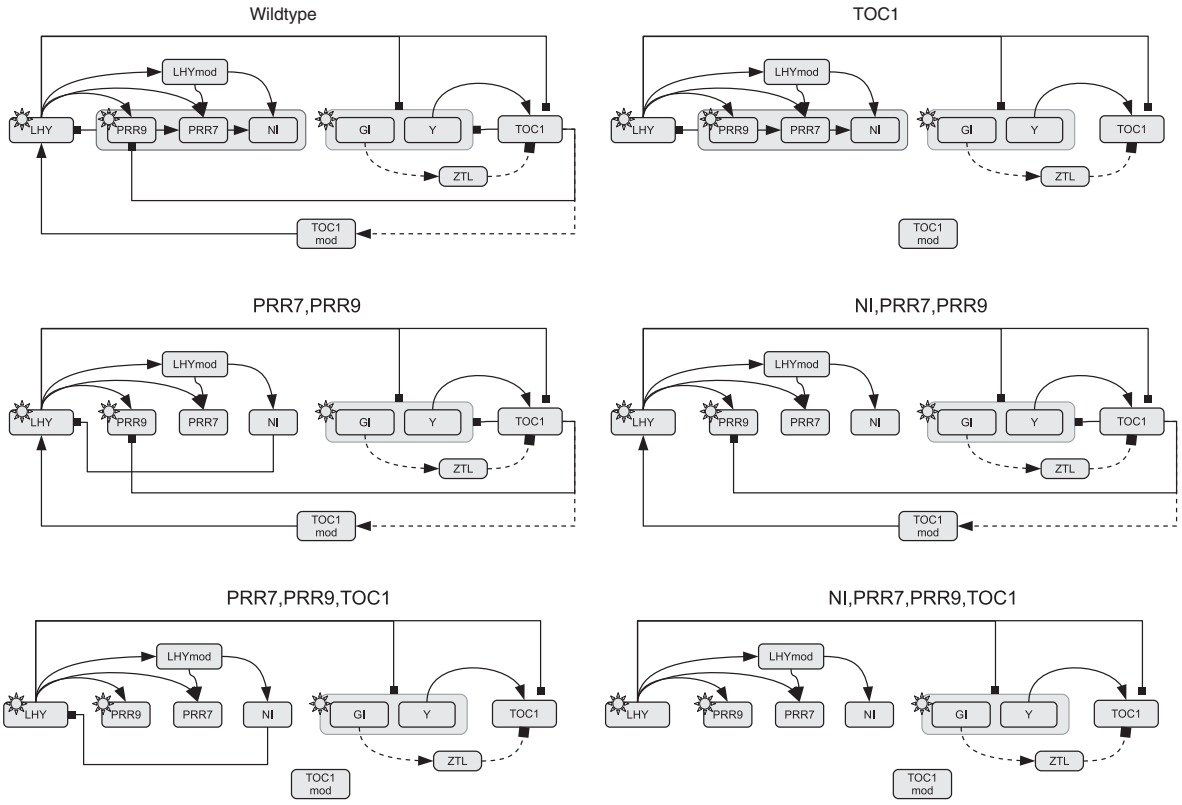

Figure 3 Model network of the circadian clock in *A. thaliana* and network modifications.
Each graph shows interactions among core circadian clock genes. Solid lines show protein-gene interactions; dashed lines show protein modifications; and the regulatory influence of light is symbolized by a sun symbol. The top left panel ("wildtype") shows the network structure published by Pokhilko et al. (2010). The remaining panels show modified network structures, corresponding to constant knockouts of the proteins shown above the corresponding network structure. Gray boxes group sets of regulators or regulated components. Arrows symbolize activations and bars inhibitions.

the circadian clock from the biological literature (Locke et al., 2005; Kolmos et al., 2009; Herrero et al., 2012; Pokhilko et al., 2010, 2012, 2013). It is important to note that, as opposed to the realistic data described in the previous subsection, we do not have a proper ground truth. Besides the fact that these models show noticeable variations, they were not obtained on the basis of proper statistical model selection, as described, e.g., by Vyshemirsky and Girolami (2008). Nevertheless, a qualitative comparison will reveal to what extent the postulated interaction features and structural network characteristics from the literature are consistent with those inferred from the data.

The data used in our study come from the EU project TiMet (2014), whose objective is the elucidation of the interaction between circadian regulation and metabolism in plants. The data consist of transcription profiles for the core clock genes from the leaves of various genetic variants of *A. thaliana*, measured with qRT-PCR. The study encompasses two wildtypes of the strains Columbia (Col-0) and Wasilewski (WS) and 5 clock mutants, namely a double knock-out *lhy/cca1* in the WS strain, a single knock-out *gi* and *toc1* in the strain Col-0, a double-knockout *prr7/prr9* in strain Col-0, and a single knock-out of *elf3*.The plants were grown in the following 3 light conditions: a diurnal cycle with 12 h light and 12 h darkness (12L/12D), an extended night with full darkness for 24 h (DD), and an extended light with constant light (LL) for 24 h. An exception is the *elf3* mutant, which was grown only in 12L/12D condition. Samples were taken every 2 h to measure mRNA concentrations. Further information on the data and the experimental protocols is available from TiMet (2014). For our study, we recorded the transcription profiles of the core clock genes that are included in the models from the literature (Pokhilko et al., 2010; Guerriero et al., 2012): *LHY, CCA1, NI (PRR5), PRR7, PRR9, TOC1, ELF3, ELF4, LUX,* and *GI*.

4 Methodological details

4.1 Preparation of realistic data

We used the Bio-PEPA framework (Ciocchetta and Hillston, 2009) to generate mRNA and protein concentration profiles with Markov jump processes. As discussed in Section 3.1, these profiles are sensitive to the choice of the unit volume parameter Ω. For values of $\Omega<10$, the concentration profiles are dominated by stochasticity, whereas for $\Omega>1000$ they become indistinguishable from the deterministic solutions of ODEs. In our study, we used a value of $\Omega=100$, as suggested by Guerriero et al. (2012), which gives the best match to the experimental qRT-PCR data, in particular with respect to the fluctuations of the qRT-PCR amplitudes. We simulated mRNA and protein concentration time series from the circadian regulatory network of Guerriero et al. (2012) and Pokhilko et al. (2010), shown in the top left panel of Figure 3, named "wildtype." In addition, we simulated mRNA and protein concentration time series from a series of modified network structures, in which various feedback loops and recurrent interactions had been removed;[14] these networks are shown in the remaining panels of Figure 3. For each of these network types we created 11 interventions, in emulation of the biological protocols of TiMet (2014) and Edwards et al. (2010). These interventions include knock-outs of the genes *GI, LHY, TOC1*, and the double knock-out of *PRR7/PRR9*. The knock-outs were simulated by down-regulating the transcription rates of the targeted genes, and replacing them by random noise, drawn from a truncated normal distribution (to ensure non-negativity of the concentrations). Again, in emulation of the biological protocols of TiMet (2014) and Edwards et al. (2010), we simulated varying photo-periods of 4, 6, 8, 12, and 18 h as well as a full dark (DD) and a full light (LL) cycle, each following a 12 h–12 h light-dark cycle entrainment phase over 5 days. For each type of intervention, concentration time series were generated to encompass a simulated epoch of 6 days, of which the first 5 days were used for entrainment. After entrainment, molecule counts of mRNA and proteins were recorded in 2-h intervals of simulated time, for 24 h, giving a total of 13 "observations." Combining these 13 observations for each intervention type yields 143 observations in total for each of the regulatory network structures shown in Figure 3. For each intervention type and sampling interval length, five independent data sets were generated; this corresponds to five independent laboratory experiments. For the results reported in this paper, the data was not log transformed. However, we compared the learning accuracy obtained from the original and the log transformed data. The results, presented in Appendix A.2, suggest that the log transformation is counter-productive. This is consistent with the fact that a log transformation leads to more complicated expressions in equation (1), as a consequence of the chain rule of differential calculus, which renders the learning task harder. To standardize the data, we followed the widely established procedure to rescale all molecule concentrations to zero mean and unit standard deviation. Two different data types were used in our evaluation procedures: complete data, which include both the mRNA and the protein concentrations, and incomplete data, in which the protein concentrations are missing and regulatory network structures have to be inferred on the basis of mRNA concentrations alone.

In summary, we generated data for six different network structures, shown in Figure 3, repeating each data generation five times independently (i.e., starting from different random number generator seeds), and using complete observations (mRNAs and proteins) versus incomplete observations (mRNA only).

4.2 Preparation of real data

The mRNA profiles for the genes *LHY, CCA1, NI, PRR7, PRR9, TOC1, ELF3, ELF4, LUX,* and *GI* were extracted from the TiMet data TiMet (2014), yielding a total of 266 samples per gene. We used the mean copy number of mRNA per cell and applied a gene wise Z-score transformation for data standardization. We did not log transform the data following the analysis in Appendix A.2. An additional binary light indicator variable with 0 for darkness and 1 for light was included to indicate the status of the experimentally controlled light condition.

14 We turned off the translation of those proteins contributing to interactions we like to surpress.

4.3 Rate estimation

Motivated by the fundamental equation of transcriptional regulation, equation (1), the machine learning and statistical models applied in our study aim to predict the rate of gene transcription from the concentrations of the putative regulators. With de novo mRNA profiling assays, the rate of transcription could in principle be measured, but these data are often not available. We therefore applied two numerical procedures to obtain the transcription rate. Appreciating that the transcription rate is just the time derivative of the mRNA concentration $x(t)$, the first approach is to approximate it by a difference quotient:

$$\frac{dx}{dt} \approx \frac{x(t+\delta t)-x(t-\delta t)}{2\delta t} \tag{55}$$

This is a straightforward procedure, and two different values for δt were used in our study: δt=2 h, henceforth referred to as the coarse gradient, and δt=24 min, henceforth referred to as the fine gradient. However, it is well known that differencing noisy time series leads to noise amplification. As an alternative procedure, we therefore used an approach based on smooth interpolation with Gaussian processes. We followed Solak et al. (2002) and exploited the fact that the derivative of a Gaussian process is a Gaussian process again; hence analytic expressions for the mean and the standard deviation of the derivative are available (Solak et al., 2002). For the covariance of the Gaussian process, we used the squared exponential kernel, which is the standard default setting in the R package `gptk` (Kalaitzis et al., 2013).

We note that motivated by equation (1), all methods included in our comparative evaluation study aim to predict the time derivative of a target gene's mRNA concentrations from either the protein (complete data) or the mRNA (incomplete data) concentrations of the putative regulators. Where a method has not been originally designed for this purpose, a few trivial modifications have to be implemented; e.g., for a dynamic method that aims to predict time-shifted target mRNA concentrations at time point t+1 from mRNA concentrations at time point t, the time shift has to be undone, and the target mRNA concentration has to be replaced by its time derivative. Motivated by equation (1), a forced link from a target gene's mRNA concentration to its time derivative is built into all regression methods to allow for mRNA degradation (represented by the linear decay term in equation (1)); this is a natural implementation of biological prior knowledge about the nature of transcriptional regulation.

4.4 Regulatory effect of the light

We note that the entity P was introduced in the circadian clock model of Guerriero et al. (2012) to model the regulatory effect of the light appropriately. In Guerriero et al. (2012) P was referred to as the "light-induced *protein*," though it was de facto employed to represent a biologically unknown light-stimulated component of the circadian clock. As the model of Guerriero et al. (2012) does not generate mRNA concentrations for P, we use the ("protein") concentration of P as potential regulator for both data scenarios. That is, in the complete data scenario we follow Guerriero et al. (2012) and think of P as a protein, while we think of P as a gene in the incomplete data scenario. Moreover, to be consistent with the model of Guerriero et al. (2012) we set the concentration of P to zero in the absence of light.[15] Whenever we infer P to be a regulator for a target gene g, we conclude that the transcription rate of g, symbolically y_g, depends on the light condition.

4.5 Gene knock-outs and mutagenesis

Both the real and the realistic data contain mutagenesis experiments with loss-of-function mutants; see Sub-Section 3. Genes that have been knocked out have to be excluded as target variables, since their values result

15 In the model equations defined by Guerriero et al. (2012) the concentration of P only appears in a product with the binary light indicator L, where the light variable L is equal to zero in the absence of light.

from external interventions and cannot be predicted *per se* from the expression status of their regulators. By treating the entire regulatory network as a union of bi-partite graphs – one target gene against all putative regulators – this exclusion becomes straightforward: each bi-partite network is inferred from only those experiments in which the target gene was *not* knocked out. Below, we provide further details on how the methods included in our comparative evaluation were applied specifically.

4.6 Method setup

As motivated in Section 4.3 we implemented the network inference algorithms, described in Section 2, with two general modifications to account for (i) mRNA degradation and (ii) mutagenesis experiments. (i) With regard to the mRNA degradation we implemented all inference methods that explicitly select regulator sets $\boldsymbol{\pi}_g$ for the target variables y_g such that the target gene's mRNA concentration, x_g, is permanently included as a member of $\boldsymbol{\pi}_g$.[16] This corresponds to the rightmost term in equation (1), and does not contribute to the target node's fan-in. (ii) For mutagenesis experiments we excluded all data points corresponding to experiments where the target gene y_g was knocked out. We note that this yields varying numbers of data points $T_g \leq T$ for each target variable y_g (g=1,…,G).

4.6.1 Graphical Gaussian models (GGM)

We used the original code of the GGM method by Schäfer and Strimmer (2005), which is implemented in the R package `GeneNet` and available from the CRAN R archive. We obtained the partial correlation matrices using function `ggm.estimate.pcor` with the `static` method and default parameter settings. From these matrices only those partial correlations were extracted that involved the target gradient response y_g. To obtain the partial correlations for the complete system, including partial correlations for all gradient responses y_g, $\forall g$, the GGM learning algorithm had to be applied repeatedly for each individual gradient response variable y_g. We treated the absolute partial correlation values as indicator for the interaction ranks.

4.6.2 Lasso, Elastic Net and Tesla

For Lasso and the Elastic Net we used the R software package `glmnet`, described by Friedman et al. (2010). We optimized the regression parameters with cyclical coordinate descent, as implemented in the `glmnet` package. The regularization parameters were selected so as to minimize the mean square cross-validation error, using a 10-fold cross-validation scheme. This was done automatically with the function `cv.glmnet()`. Absolute values of non-zero regression coefficients were recorded and used for ranking molecular interactions.

Tesla was run with a linear regression implementation in Matlab. The regression parameters were optimized with convex programming, using the `CVX MATLAB` package.[17] A 10-fold cross-validation scheme was applied to optimize the regularization parameters, minimizing the mean square cross-validation error. Tesla requires the prior specification of permissible change-points. We selected light as the primary segmentation criterion, and grouped measurements obtained under the same light condition (light versus darkness) together. This gives, for each gene, two different segments with potentially different regression parameters. The absolute values of the non-zero regression coefficients were recorded for both segments, and their averages were used for ranking the molecular interactions.

16 For the Bayesian methods this can be enforced by setting the prior $P(\boldsymbol{\pi}_g)$ to zero for all $\boldsymbol{\pi}_g$ with $x_g \notin \boldsymbol{\pi}_g$.
17 Matlab software for *Disciplined Convex Programming:* http://cvxr.com/cvx/.

4.6.3 Hierarchical Bayesian regression (HBR)

The MCMC simulations for the Bayesian regression methods, with and without multiple change-points, as described in Sections 2.5 and 2.6, were run for 20,000 iterations each, with a burn-in period of 10,000 iterations discarded. This choice gave satisfactory convergence diagnostics, based on correlation scatter plots and Gelman-Rubin potential scale reduction factors (Gelman and Rubin, 1992; Brooks and Gelman, 1999). Marginal posterior probabilities of molecular interactions were obtained from the MCMC trajectories, estimated from the relative frequency of inclusion of the corresponding edges in the sampled models.

4.6.4 Sparse Bayesian regression with automatic relevance determination (ARD-SBR)

For the sparse Bayesian regression approach with automatic relevance determination (SBR-ARD) we used the MATLAB implementation from the supplementary material of Rogers and Girolami (2005). We used the default settings both for the hyperparameters and the maximal number of iterations for the marginal likelihood maximization. We note that the method in Rogers and Girolami (2005) is a slightly modified version of the fast marginal likelihood algorithm from Tipping et al. (2003); for the technical details we refer the reader to the supplementary material of Rogers and Girolami (2005).

4.6.5 Bayesian splines autoregression (BSA)

We used the MATLAB programs provided with the supplementary material of Morrissey et al. (2011), with the following modification: for the target genes, we replaced the future gene expression values by the estimate of the time derivatives, y_g, as discussed at the end of Section 2.8. This implementation is particularly straightforward for the gene-specific hyperparameters, corresponding to equations (2.8–2.9) in Morrissey et al. (2011), which we elected to use, as no difference between gene-specific and global hyperparameters was found by Morrissey et al. (2011). For the other model options, including the order of the splines, the number of knots, and the hyperparameters of the Bayesian model, we used the default settings in the MATLAB programs; note that they had been applied by Morrissey et al. (2011) to data from a very similar model (also related to circadian regulation in Arabidopsis). For the MCMC simulations, we proceeded in the same way as for the Bayesian methods, applying standard convergence diagnostics based on potential scale reduction factors.

4.6.6 State-space models (SSM)

In its multivariate formulation, the SSM methods described in Section 2.9 can neither deal with target-specific potential regulator sets nor with the required target-specific exclusion of certain data in relation with mutagenesis experiments (note that mRNA concentrations of knock-out genes are the result of external interventions and cannot be predicted from within the system). However, this can be easily rectified by implementing a separate SSM for each target variable y_g, which ensures a fair comparison with the other methods. For approximate inference with the variational Bayesian EM-algorithm, we used the Matlab implementation from Beal (2003). We used the default parameter settings and varied the number of hidden nodes (i.e., the dimensionality of the vector $\mathbf{h}$) from n=1 to n=8.[18] We trained two target-specific SSMs for each n=1,…,8, starting from two different random initializations, i.e., 7·2·8=112 target-specific SSMs in total. Except for low values of n ($n \leq 2$), where we observed slightly deteriorated AUROC values for the incomplete data, we obtained very stable network predictions in terms of the posterior expectation of the interaction matrix

18 Note that the maximal number of hidden nodes n is restricted by the number of regulators, G_g. In our simulation study we analyzed various data sets, and we employed the lowest G_g as an upper bound on the number of hidden nodes n.

elements $(\mathbf{CB}+\mathbf{D})_{g,i}$ (g=1,…,G and i=1,…,G_g). Throughout the paper we therefore only report the network reconstruction results that we obtained with n=8 hidden nodes, noting that almost identical results could have been obtained for n=3,…,7.

4.6.7 Gaussian process (GP)

For the Gaussian process approach described in Section 2.10, we used the implementation in the `GP4GRN` software package, developed by Äijö and Lähdesmäki (2009). This software computes, for each target gene, the posterior probabilities of all potential sets of regulators. The posterior probabilities for individual molecular interactions are then obtained by marginalization, summing the posterior probabilities of all configurations of regulators that include the molecular interaction in question, as shown in equation (40). The hyperparameters $\boldsymbol{\theta}$, σ^2 in equation (40) were optimized with the Polack-Ribiere conjugate gradient method (Rasmussen and Williams, 2006) to maximize the marginal likelihood of equation (38). Following Äijö and Lähdesmäki (2009), the hyperparameters $\mathbf{b}$, σ_b^2 were set fixed. We chose the same values as suggested by Äijö and Lähdesmäki (2009). In addition, we tried a selection of randomly perturbed values, and computed the average performance. We then selected whichever of these two alternatives achieved the higher AUROC score.

4.6.8 Mutual information methods (ARACNE)

The application of the mutual information approach was conducted with the ARACNE method. We used the R package `minet` (Meyer et al., 2008) from the Bioconductor package, which includes a function to build a mutual information matrix (`build.mim`) together with the actual ARACNE implementation (`aracne`). We used the default settings with the Spearman's correlation estimator, and no discretization for building the mutual information matrix. This matrix is passed to function `aracne`, which in turn produces a weighted adjacency matrix by removing the weakest links given a triplet of links subject to a threshold, which we kept at the default value. The relevant links that involve the target gradient response y_g were extracted from the adjacency matrix and directly used as indicator for the interaction ranking. To construct the full network, the whole procedure was repeated for each target gene g.

4.6.9 Mixture Bayesian network models (MBN)

For the mixture Bayesian network (MBN) approach we applied the implementation of the EM-algorithm for Gaussian mixture models from the "Pattern Analysis Toolbox" by I.T. Nabney; this Matlab toolbox has been made available as supplementary material of Nabney (2002). As the EM-algorithm is a greedy optimization technique that converges to the nearest (local) maximum of the likelihood, we repeated the application 10 times, starting from different initializations.[19] This yields H=10 regulator sets $\boldsymbol{\pi}_g^{(1)},\ldots,\boldsymbol{\pi}_g^{(10)}$ for each target variable y_g (g=1,…,G). In imitation of the Bayesian approach we use the fraction of regulator sets that obtain the regulator x_j^g to rank the regulatory interactions $x_j^g \to y_g$ (g=1,…,G and j=1,…,G_g).

19 In our study we initialized the EM-algorithm with allocations obtained by the k-means cluster algorithm. Thereby the initial $\mathcal{K}_g$ centers of the k-means algorithms were sampled from a multivariate Gaussian $N(\boldsymbol{\mu}, \mathbf{I})$ distribution, where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\mu}$ is a random expectation vector with entries sampled independently from continuous uniform distributions on the interval [–1, +1]. To avoid that the EM-algorithm is initialized with allocations that possess unoccupied (empty) mixture components, we re-sampled the initial centers and re-ran the k-means algorithm whenever we obtained k-means outputs with empty components.

4.6.10 Gaussian Bayesian networks (BGe)

For the Gaussian Bayesian network model with the BGe scoring metric the prior distribution of the unknown parameters is assumed to be a Gaussian-Wishart distribution with hyperparameters α, $\mathbf{T}_0$, ν, and $\boldsymbol{\mu}_0$. In the absence of any genuine prior knowledge about the regulatory interactions we set the parameter matrix of the Wishart prior to the identity matrix, symbolically $\mathbf{T}_0=\mathbf{I}$, and the mean vector of the Gaussian prior to the zero vector, symbolically: $\boldsymbol{\mu}_0=\mathbf{0}$.[20] The scalar hyperparameters α and ν, which can be interpreted as equivalent prior sample sizes (see Geiger and Heckerman, 1994), were set to $\alpha=G_g+4$ and $\nu=1$. That is, we set the equivalent prior sample sizes as uniformative as possible subject to the regulatory conditions discussed by Geiger and Heckerman (1994). We imposed a maximal fan-in restriction of $\mathcal{F}=3$, which renders the computation of the marginal interaction posterior probabilities in equation (54) computationally tractable.

4.7 Network inference scoring scheme

All the methods described in Section 2 provide a means by which interactions between genes and proteins can be ranked in terms of their significance or influence. If the true network is known, this ranking defines the Receiver Operating Characteristic (ROC) curve, where for all possible threshold values, the sensitivity or recall is plotted against the complementary specificity.[21] By numerical integration we then obtain the area under the curve (AUROC) as a global measure of network reconstruction accuracy, where larger values indicate a better performance, starting from AUROC=0.5 to indicate random expectation, to AUROC=1 for perfect network reconstruction. There have been suggestions that precision-recall curves indicate differences in network reconstruction performance more clearly than ROC curves (Davies and Goadrich, 2006). While this is true for large, genome-wide networks, our own previous work has indicated that for the network complexity of interest in our study, indicated in Figure 3 and the publications on circadian clock modeling, Locke et al. (2005) and Pokhilko et al. (2010, 2012, 2013), the differences between the two scoring schemes are negligible (Grzegorczyk and Husmeier, 2013). We therefore evaluate the performance of the methods described in Section 2 with AUROC scores, due to their more straightforward statistical interpretation (Hanley and McNeil, 1982).

4.8 ANOVA

For our evaluation, we were running hundreds of simulations for a variety of different settings, related to the observation status of the molecular components (mRNA only versus mRNAs and proteins), the method for derivative (rate) estimation (described in Section 4.3), the regulatory network structure (shown in Figure 3), and the method applied for learning this structure from data (reviewed in Section 2). The results, depicted, e.g., in Figures 11 and 15, are complex and elude clearly discernible patterns and trends. In order to disentangle the different factors, and in particular distinguish the effect of the model from the other confounding factors, we adopted the DELVE evaluation procedure for comparative assessment of classification and regression methods in Machine Learning (Rasmussen, 1996; Rasmussen et al., 1996) and set up a multi-way analysis of variance (ANOVA) scheme (e.g., Brandt, 1999).

Let y_{ognmk} denote the AUROC score obtained for observability status o, gradient computation g, network topology n, network reconstruction method m, and data instantiation k. The range of these index parameters is as follows: $o\in\{0, 1\}$, where $o=0$ indicates partial (mRNAs only) and $o=1$ complete (mRNAs and proteins)

20 Loosely speaking, this setting ($\boldsymbol{\mu}_0=0$ and $\mathbf{T}_0=\mathbf{I}$) reflects our "prior belief" that all domain variables, i.e., the potential regulators and the target variable, are i.i.d. standard normally distributed.

21 The *sensitivity* is the proportion of true interactions that have been detected, the *specificity* is the proportion of non-interactions that have been avoided.

observation; $g \in \{0, 1, 2\}$, where $g=0$ denotes coarse gradient, $g=1$ fine gradient, and $g=2$ gradient from a smooth interpolant; $m \in \{0, 1, 2, 3, 4, 5\}$, where $m=0$ represents "wildtype" (the published network topology), and $m \neq 0$ the five network modifications shown in Figure 3; $n \in \{0, 1, 2, \ldots, 14\}$, for the 15 network reconstruction methods discussed in Section 2 (and shown below in Figure 6), and $k \in \{0, 1, 2, 3, 4\}$ for five different data instantiations. We model the AUROC scores with the following ANOVA approach:

$$y_{ognmk} = O_o + G_g + N_n + M_m + \varepsilon_{ognmk} \tag{56}$$

where $\varepsilon_{ognmk} \sim N(0, \sigma^2)$ is zero-mean white additive Gaussian noise, and O_o, G_g, N_n, and M_m are main effects associated with observation status, gradient computation, network topology, and network reconstruction method, respectively. As a sanity check, we carried out standard residual analysis; this did not indicate any violation of the model assumptions, as discussed in Appendix A.1.

5 Results

Our Results section can be divided into three parts. In the first part, which covers Sections 5.1–5.3, we address questions related to the application of the models: how to set the regularization parameters for the sparse regression methods, and how to set the parameter and structure prior for the Bayesian regression models. In the second part, we address the main questions of our study: How do the different methods compare with respect to the accuracy of the network reconstruction? What is the effect of missing protein concentrations? What is the effect of the network topology? What is the best way to compute the transcription rates? What is the effect of change-points, both for the light phase and the rates? This part covers Sections 5.4–5.9 and, like the previous part, is based on the realistic data described in Section 3.1. The final part, Section 5.10, describes the application to the real data from Section 3.2.

5.1 Comparison of different methods for setting the Lasso penalty parameter

The sparse regression methods Lasso and Elastic Nets require the selection of a regularization parameter λ, which trades off the strength of the L1 or L1/L2 penalty term against the data misfit term. We have compared three different procedures: 10-fold cross-validation, 10-fold cross-validation with the correction suggested by Hastie et al. (2001), and BIC. The objective of 10-fold cross-validation is to select the regularization parameter that minimizes the average signal reconstruction error on held-out data. Hastie et al. (2001) suggested using a larger value as follows: plot the cross-validation error as a function of λ, then select the largest value of λ for which the cross-validation error is within 1 standard deviation of the minimum cross-validation error (Hastie et al., 2001). The rationale is that Lasso is biased (Murphy, 2012) and that the optimal value of λ chosen by cross-validation is optimal in terms of predictive (signal reconstruction) rather than explanatory (network connectivity) performance. Hastie et al. (2001) suggested this correction as a heuristic scheme to improve explanatory performance. The motivation for using BIC is to avoid the computational costs of a cross-validation scheme. We compared the Lasso with these three procedures on our simulated data described in Section 3.1, which includes several network types (as shown in Figure 3), incomplete (mRNA only) and complete (mRNA and protein) data, and coarse and fine gradients (Section 4.3). Figure 4 shows a Bland-Altman plot, where the difference between the AUROC scores are plotted against the mean AUROC scores. A visual inspection suggests that standard minimum cross-validation achieves slightly higher AUROC scores on average than the other two methods. A paired t-test confirms that standard cross-validation performs significantly better than the procedure proposed in Hastie et al. (2001) (p-value of 0.0004), and weakly outperforms BIC (p-value of 0.10). The standard minimum cross-validation approach thus performs overall best and was used for the further investigations.

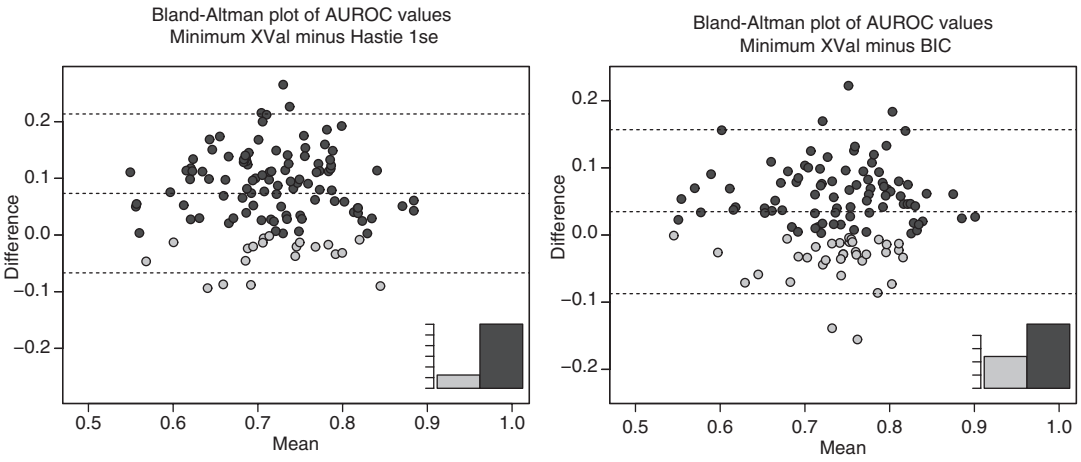

Figure 4 Bland-Altman plot of AUROC scores comparing different selection procedures for the regularization parameter of Lasso. The plots print the difference between the AUROC scores (vertical axis) against their mean (horizontal axis). Left panel: Comparison between standard minimum cross-validation and the procedure proposed by Hastie et al. (2001). Right panel: Comparison between minimum cross-validation and BIC. For details, see Section 5.1. Positive values (dark gray) indicate higher AUROC scores for the standard cross-validation procedure. Negative values (light gray) indicate higher scores for the alternative procedure (Hastie et al. or BIC). The inlet histograms in the bottom right corner show the relative frequencies of positive (dark gray) and negative (light gray) scores.

5.2 Influence of the structure prior for hierarchical Bayesian regression models

We tested the Bayesian regression models, described in Sections 2.5 and 2.6, with two different prior distributions on the network structure: a uniform distribution and a truncated Poisson distribution for $P(\boldsymbol{\pi}_g)$. The Poisson prior has mean κ and a maximal cardinality matching that of the uniform prior, i.e., $|\boldsymbol{\pi}_g| \leq 3$: $P(\boldsymbol{\pi}_g \mid \kappa) \propto \frac{\kappa^{|\boldsymbol{\pi}_g|}}{|\boldsymbol{\pi}_g|!} I(|\boldsymbol{\pi}_g| \leq 3)$, where κ is sampled from a vague conjugate prior with a gamma distribution $P(\kappa) = \mathcal{G}a(0.5, 1)$, following Lèbre et al. (2010). We tested the Bayesian regression model with both priors on the simulated data, described in Section 3.1, and recorded the AUROC scores. The left panel in Figure 5 shows a Bland-Altman plot with the pair-wise differences of the AUROC values (vertical axis) plotted against the mean AUROC score (horizontal axis). The plot shows no noticeable difference between the two priors, and a paired t-test with a p-value of 0.17 indicates no significant difference. We decided to use the uniform prior for all further investigations.

5.3 Influence of the parameter prior for hierarchical Bayesian regression models

We compared two different priors on the regression parameters of the Bayesian regression model described in Sections 2.5 and 2.6: the so-called ridge regression prior of equation (21), and the g-prior. The latter is widely used in the statistics literature, see e.g., Andrieu and Doucet (1999) and Marin and Robert (2007), and effectively replaces the diagonal matrix in equation (21) by an outer product of the design matrix; see Marin and Robert (2007) for details. We carried out a comparative evaluation on the realistic data from Section 3.1. The right panel in Figure 5 shows a Bland-Altman plot of the pairwise differences in the AUROC scores. There is a slight shift to positive values, indicating that, overall, the ridge regression prior achieves a better performance. This difference was found to be significant, with a paired t-test giving a p-value of $2.6e{-}19$. We therefore used the ridge regression prior of equation (21) throughout our study.

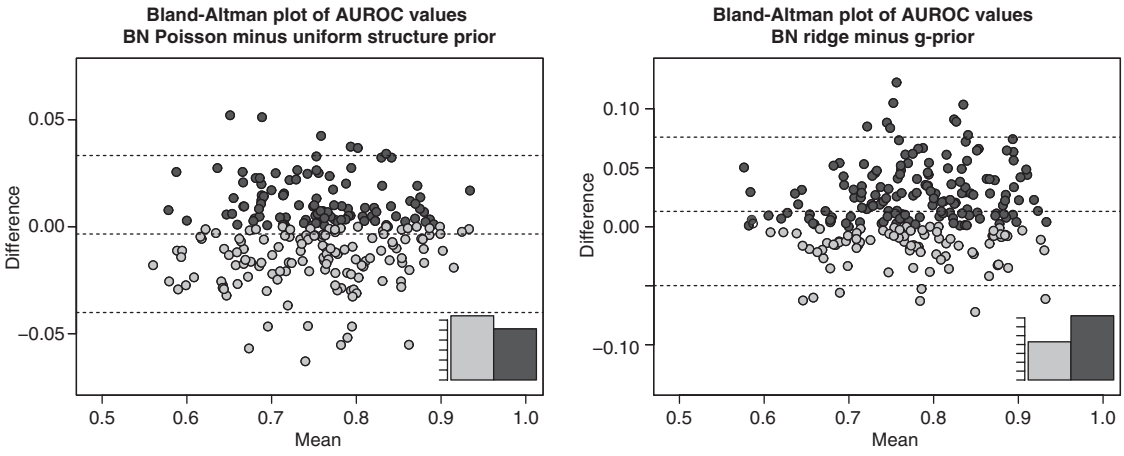

Figure 5 Dependence of Bayesian regression on the structure (left panel) and parameter (right panel) prior.
The graphs show Bland-Altman plots, which plot the difference between two AUROC scores (vertical axis) against their mean (horizontal axis). *Le*ft panel: Difference between the uniform and the Poisson structure prior. Positive values (dark gray dots) indicate AUROC scores in favor of the uniform prior, negative values (light gray dots) indicate AUROC scores in favor of the truncated Poisson prior. Right panel: Difference between ridge regression prior and *g*-prior. Positive values (dark gray dots) indicate better performance of the ridge regression prior, negative values (light gray dots) indicate better performance of the g-prior. The inlet histograms in the bottom right corner show the relative frequencies of positive (dark gray) and negative (light gray) scores.

5.4 Comparison between the methods

A main objective of our study is a systematic comparative performance evaluation of the models reviewed in Section 2. These models were applied to the different data described in Section 3, different observabilities (proteins and mRNAs versus mRNAs only), different gradient computations (Section 4.3), and different network topologies (as shown in Figure 3). Figure 15 shows the distributions of AUROC scores obtained in our study. The scores vary considerably, depending on the different factors, and consistent trends and clear patterns are not easily discernible. To enable a clearer interpretation we adopted the ANOVA method described in Section 4.8. The quantity of interest is M_m – the main effect of the network reconstruction method, which is plotted in Figure 6. Our study suggests that with the exception of MBN and ARACNE, which show a significantly worse performance, all methods achieve a performance in the range of AUROC scores between 0.7 and 0.8. This is significantly better than random expectation, but considerably worse than perfect network reconstruction. The best performance is achieved with BGe and HBR. A somewhat surprising finding is that within the group of Bayesian regression models, no performance improvement is achieved by including change-points to indicate the light phase (HBR-light), change-points in the amplitude to model Michaelis-Menten non-linearities (HBR-cps), or non-linear (inverse and quadratic) terms (HBR-nl). In fact, the simple linear Bayesian regression model with no change-points (HBR) achieves the best performance of all the methods included in the comparison. This seems counter-intuitive, given that light has a clear influence on circadian regulation, and the processes of the underlying Michaelis-Menten kinetics are intrinsically non-linear. We discuss the reason for this behaviour in Section 6, where we also provide explanations for the poorer performance of some of the alternative models.

5.5 Influence of rate estimation

The mathematical formulation of chemical kinetics, e.g., based on mass action or Michaelis-Menten kinetics, as in the present study, predicts the rate of mRNA transcription as a function of the concentrations of the regulating proteins. Ideally, this rate would be measured, which could in principle be effected with *de novo*

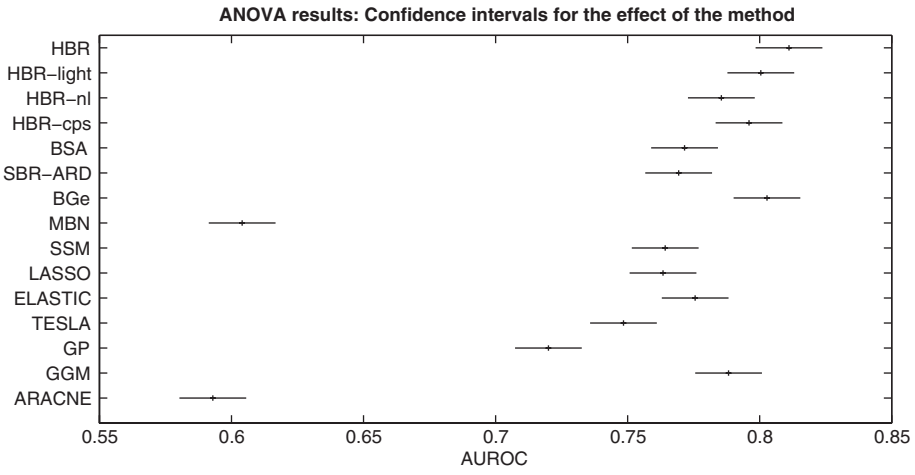

Figure 6 Comparion between different network reconstruction methods.
The figure shows confidence intervals for the group means associated with the main effect for the network reconstruction method from the ANOVA analysis in equation (56). The horizontal axis shows the AUROC score. The vertical axis represents the different methods described in Section 2. The labels on the vertical axis refer to the methods described in Section 2, using the same abbreviations as in the subtitle headers.

mRNA assays. These assays are not always available, though; so in the present study, we estimated the time derivatives of mRNA concentrations directly from the mRNA concentration time courses as a proxy. We compared three different approaches. In the first study, we approximated the time derivatives by finite difference quotients from the low frequency time series, where observations were taken every 2 h. This corresponds to equation (55) with δt=2 h, and we refer to it as the coarse gradient. In the second study, we repeated the same procedure on high-frequency data, where measurements were taken every 24 min. This corresponds to equation (55) with δt=24 min, and we refer to this as the fine gradient. High frequency data with such short time intervals are rarely available in practice, though. So as an alternative, we applied a Gaussian process smoothing approach described in Section 4.3. The results are shown in Figure 7. It can be observed that the fine gradient achieves an improvement on the coarse gradient, which is consistent with expectation. However, our study also allows a quantification of this improvement, which is in the order of Δ AUROC=0.02 on average. Interestingly, our study suggests that gradient computation in combination with smooth interpolation using Gaussian processes achieves an even more substantial improvement of about Δ AUROC=0.03. This indicates that intelligent data preprocessing leads to a better boost in predictive performance than blindly carrying out additional experiments.

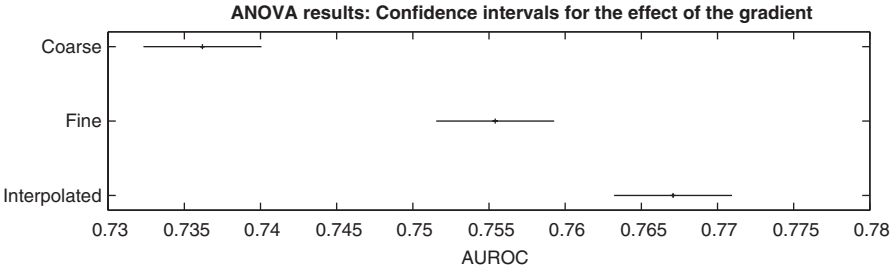

Figure 7 Influence of rate estimation.
The figure shows confidence intervals for the group means associated with the main effect for the dependence of the performance (AUROC score) on the rate estimation method, based on the ANOVA analysis of equation (56). The horizontal axis shows the AUROC score. The vertical axis represents the three rate estimation methods, as described in Section 4.3: coarse gradient (top), fine gradient (middle), and gradient from smooth interpolant (bottom).

5.6 Influence of missing protein concentrations

We have carried out the simulations for two types of data: complete observation, where both protein and mRNA concentrations are available, and incomplete observation, where protein concentrations are missing. The results are shown in Figure 8. The network reconstruction accuracy based on complete observations is slightly better than that from incomplete observations. The important new contribution of our study is to objectively assess the difference in performance, profiled over different network topologies, different ways of preprocessing the data, and different statistics and machine learning methods. This has been effected with the ANOVA approach described in Section 4.8, which quantifies the effect of missing protein concentrations as leading to a deterioration of only $\Delta AUROC = 0.002 \pm 0.003$. Hence, our study leads to the counter-intuitive finding that the difference in performance is *not* significant. We provide a discussion in Section 6.

5.7 Influence of network topology and feedback loops

An important aspect of our study is the investigation of how the network reconstruction accuracy depends on the connectivity of the network topology and the proportion of recurrent connections. To this end we have successively pruned feedback interactions, as shown in Figure 3. Figure 9 suggests that there is a notice-

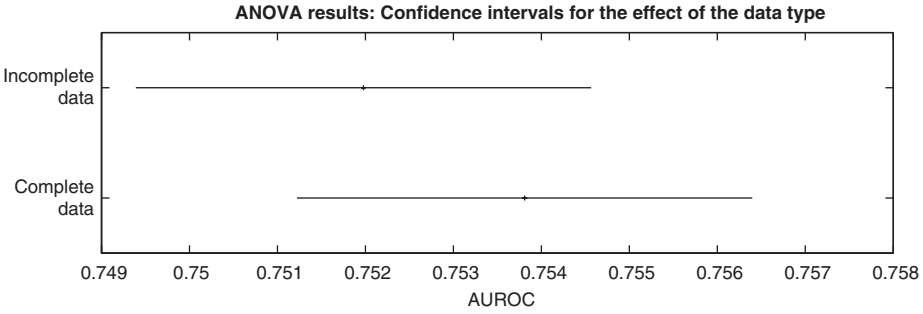

Figure 8 Influence of incomplete observations.
The figure shows confidence intervals for the group means associated with the main effect for the observation status based on the ANOVA analysis of equation (56), comparing complete data observations of both protein and mRNA concentrations versus the incomplete data that includes mRNA observations only. The horizontal axis represents AUROC scores.

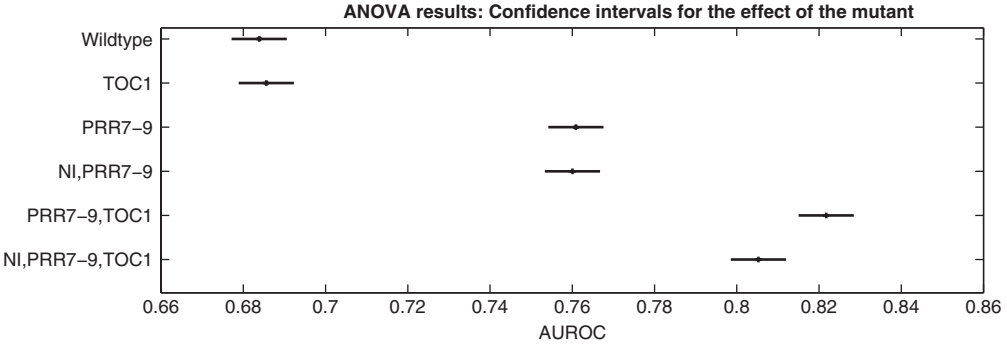

Figure 9 Influence of network structure.
The figure shows confidence intervals for the group means associated with the main effect for the network structure, based on the ANOVA analysis of equation (56), for the wildtype network published by Pokhilko et al. (2010), and the five modified structures shown in Figure 3 (using the same labels on the vertical axis as used in Figure 3). As one descends from the top to the bottom on the vertical axis, the network structures become sparser, with feedback loops increasingly being pruned. The horizontal axis represents AUROC scores.

able pattern, with less recurrent and sparser network structures appearing to be easier to learn and leading to higher AUROC scores. While this confirms a known and intuitively plausible trend, our study allows an objective quantification of the difference in performance, which has been found to amount to $\Delta AUROC$=0.14 between the most and least recurrent structures.

5.8 Influence of change-points to indicate the light phase

We tested whether a segmentation of the data into day and night phase affects the learning performance, motivated by the hypothesis that regulation in light and dark may differ and should be modeled with two separate sets of regression parameters. To this end, light information in our realistic studies described in Section 3.1 was used to assign each sample in time to a light or dark phase. We extended the range of methods to include a Lasso variant that supports change-points (Tesla, described in Section 2.4), and a non-homogeneous hierarchical Bayesian regression model, described in Section 2.6.1. Simulation experiments were conducted for incomplete (mRNA only) and complete (mRNA and proteins) data, as well as for coarse and fine response gradients, as described in Section 4.3. Figure 10 shows the distribution of pairwise differences between a method without change-points and the corresponding change-point method (Lasso versus Tesla and homogeneous Bayesian regression versus non-homogeneous Bayesian regression). The somewhat counter-intuitive finding is that for complete data (protein and mRNA concentrations, in gray boxes), the inclusion of change-points leads to a deterioration of the AUROC scores for most of the network structures. We will discuss this observation in Section 6.

5.9 Effect of change-points on the response variable

We studied the effect of segmenting the domain of the response variable (i.e., the rate, that is the time derivative of the mRNA concentration) with multiple change-points. The objective is to approximate the non-linearity of the Michaelis-Menten response with a piece-wise linear model. To this end, we applied the non-homogeneous hierarchical Bayesian regression model, described in Section 2.6.2, with different settings of the maximum

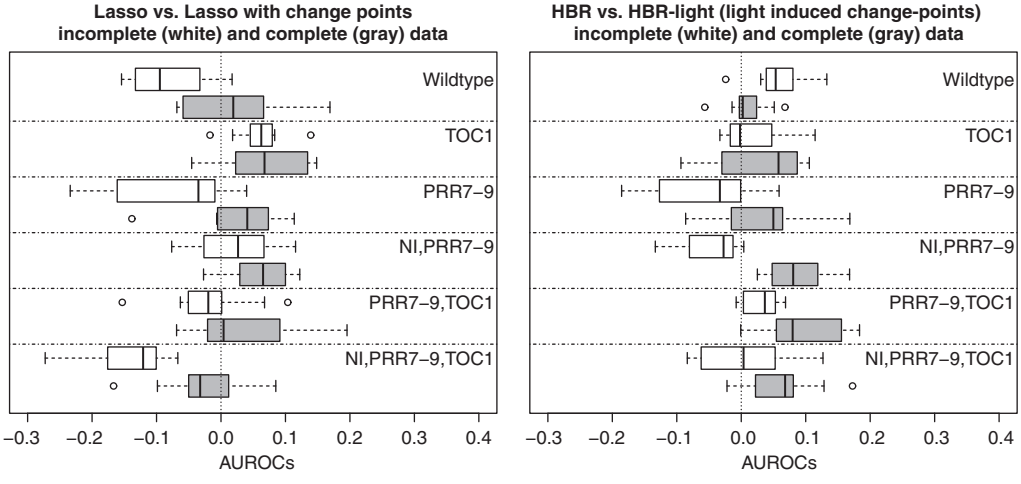

Figure 10 Dependence of the network reconstruction on light/dark phase segmentation for Lasso and Bayesian regression. The figure shows the distribution of pairwise AUROC differences for Lasso versus Tesla ($AUROC_{Lasso}$–$AUROC_{Tesla}$; left panel) and homogeneous Bayesian regression versus non-homogeneous Bayesian regression with light induced change-points ($AUROC_{withoutcps}$–$AUROC_{withcps}$). The distributions are over all network topologies (as shown in Figure 3), numerical replications, and coarse and fine gradients, as described in Section 4.3. Gray shading: complete data with protein concentrations as predictor for the target mRNA gradients. White shading: incomplete data with mRNA concentrations as predictor for the target mRNA gradients.

number of change-points. The evaluation was extended over all network topologies (shown in Figure 3), incomplete (mRNA only) and complete (mRNA and protein concentrations) data, and different gradient resolutions (coarse versus fine; see Section 4.3). The results are shown in Figure 11. The somewhat counter-intuitive finding is that the network reconstruction performance tends to deteriorate as more change-points are allowed, suggesting that despite the non-linear nature of the underlying Michaelis-Menten kinetics, imbuing the model the non-linear modeling flexibility is counter-productive. This trend is slightly stronger for complete (mRNAs and proteins) than for incomplete data (mRNAs only). We provide an explanation in Section 6.

5.10 Circadian regulation network in *Arabidopsis thaliana*

Figure 12 shows the network learned from the TiMet data, and six hypothetical networks published by Locke et al. (2006), Kolmos et al. (2009), Herrero et al. (2012), and Pokhilko et al. (2010, 2012, 2013). Solid lines show

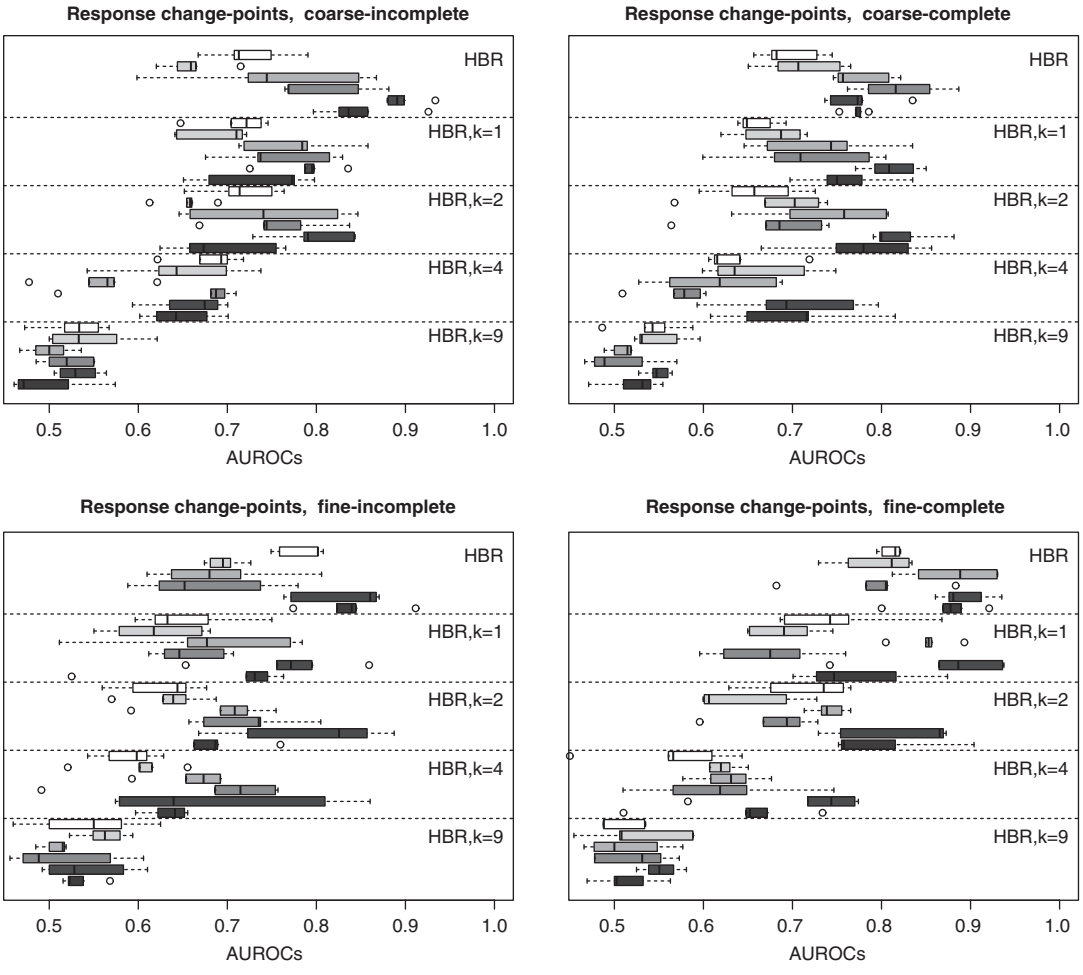

Figure 11 Dependence of the network reconstruction accuracy on the change-points for response segmentation.
The figure contains four panels. Left panels: Incomplete data, which only include mRNA concentrations. Right panels: Complete data, which include mRNA and protein concentrations. Two different gradient computations were applied, as described in Section 4.3. Top panels: coarse gradients. Bottom panels: fine gradients. Each panel contains five subpanels for five different variants of the hierarchical Bayesian regression model, described in Sections 2.5–2.6: homogeneous Bayesian regression model without change-points, and non-homogeneous Bayesian regression model with k=1, 2, 4 and 9 change-points. Each subpanel shows the distribution of AUROC scores for the six different network topologies in Figure 3, with increasing network sparsity from top to bottom.

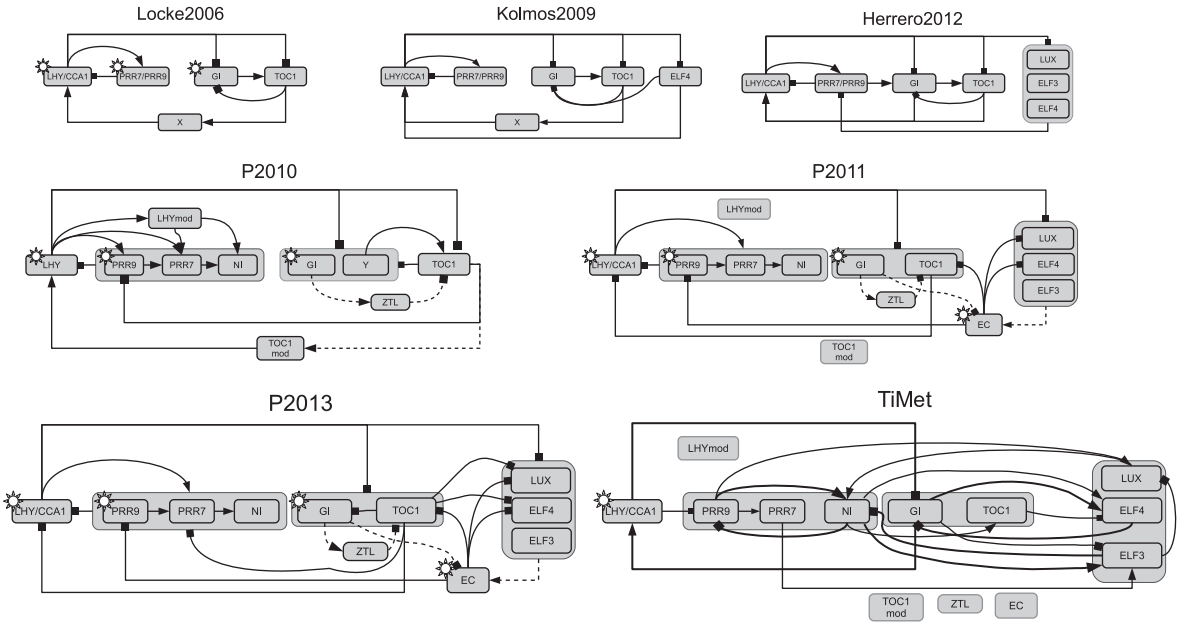

Figure 12 Hypothetical circadian clock networks from the literature, and inferred from the TiMet gene expression data. All panels except for the bottom right show hypothetical networks from the literature: Locke2006 (Locke et al., 2006), Kolmos2009 (Kolmos et al., 2009), Herrero2012 (Herrero et al., 2012), P2010 (Pokhilko et al., 2010), P2011 (Pokhilko et al., 2012), and P2013 (Pokhilko et al., 2013). The bottom right panel (TiMet) displays the reconstructed network from the TiMet data, described in Section 3.2, using the hierarchical Bayesian regression model from Section 2.5. Gene interactions are shown by black lines; protein interactions are shown by dashed lines; an arrow head symbolizes activation and a bar head inhibition; regulation by light is represented by a sun symbol. The interactions in the reconstructed network were obtained from their estimated posterior probabilities. Those above the selected threshold of 0.95 were included in the interaction network; for the light influence see the main text.

transcriptional regulation, dashed lines represent protein complex formation. The latter cannot be learned from transcriptional data and are thus systematically missing. This explains, for instance, why ZTL and EC are detached from the remaining network. The same applies to the modified proteins TOC1-mod and LHY-mod. Various features of the published networks are reproduced, though, like the acute light response in the transcription of *LHY* and *CCA1*, the activation of *PRR7* by *PRR9*, the inhibition of *GI* by *LHY/CCA1*, and the inhibition of *ELF4* by *TOC1*, which can be found in the network P2013. Various features are similar to the published networks. In the reconstructed network, *NI* is directly activated by *PRR9*, while in the published networks, the activation is indirect, via *PRR7*. The positive feedback loop from the so-called evening genes to the morning genes consists of an activation of *LHY/CCA1* by *GI*. The nature of this feedback loop (activation) is consistent with (Locke et al., 2006; Kolmos et al., 2009; Pokhilko et al., 2010). In these publications, the regulatory influence is caused by *TOC1* rather than *GI*, but these two genes are "neighbors" in the published networks (meaning: regulating each other, and exhibiting similar expression profiles). One of the morning loop genes (*NI*) is predicted to be inhibited by *ELF3*. This is consistent with Pokhilko et al. (2012, 2013), although in these publications, the interaction is indirect (via EC) and affects a neighboring target gene (*PRR9*). As mentioned above, it is intrinsically unfeasible to learn post-transcriptional processes, like protein complex formation, from transcriptional data alone; so it is no surprise to see that the protein complex EC is detached from the remaining network. It is particularly interesting to note that a key network motif repeatedly found in the reconstructed network concurs with the published networks. This is the two-node feedback motif in which a gene is the activator of its own inhibitor. This structure is particularly clearly seen in Locke et al. (2006), where it occurs three times: within the group of morning genes (*LHY/CCA1* activating *PRR7/PRR9*, *PRR7/PRR9* inhibiting *LHY/CCA1*), within the group of evening genes (*GI* activating *TOC1*, *TOC1* inhibiting *GI*), and between the morning and evening genes (*LHY/CCA1* inhibiting *TOC1*, *TOC1* activating *LHY/CCA1*). These three feedback mechanisms exist in the reconstructed network also and are highlighted with thick

lines (see TiMet network in Figure 12), involving neighboring nodes in the same three gene groups: morning genes (*PRR9* activating *NI*, *NI* inhibiting *PRR9*), evening genes (*GI* activating *ELF4*, *ELF4* inhibiting *GI*), and between morning and evening genes (*GI* activating *LHY/CCA1*, *LHY/CCA1* inhibiting *GI*, *NI* activating *ELF3*, *ELF3* inhibiting *NI*). This suggests that, despite deviations in the detailed mechanisms, the key topological features of the published networks have been successfully reconstructed. Finally, we attempted to learn the light influence marked with a sun symbol in the TiMet network in Figure 12 by allowing light as an additional variable. We correctly recovered a high probabability (0.83) link to *LHY/CCA1* but failed to observe any other significant occurrences. It was noted in Section 4.1 that the light influence on mRNA transcription is typically modulated by light sensitive proteins. Since the TiMet data lack any such protein observations, we have to assume that the light is not learned efficiently.

6 Discussion

The previous section has presented the results from our comparative evaluation study. Most of the patterns that we have found are clear and intuitive; the value of our study consists in the objective quantification of these trends. There are a few findings that are peculiar, though. Figure 6 suggests that it is counter-productive to include non-linear terms in the Bayesian regression model. Given that the true underlying dynamics are, in fact, non-linear, why does the inclusion of these effects deteriorate the model performance? Figure 11 suggests that an increasing number of change-points for the response segmentation in the non-homogeneous Bayesian regression model deteriorates the network reconstruction. However, more change-points give more non-linear modeling flexibility. Given that the true underlying dynamics are non-linear, why is that a disadvantage? How can we understand that the network reconstruction accuracy does not improve significantly when including protein concentrations in addition to just mRNA concentrations, as suggested by Figure 8. Figure 10 shows the effect of segmenting the data into a light and a dark phase. How can we understand that this segmentation deteriorates the network reconstruction accuracy for complete (mRNA plus protein) data? Finally, Gaussian processes are widely appreciated as a powerful modeling paradigm. So how can we explain their comparatively poor performance (see Figure 6)? In what follows, we will provide an explanation of these effects.

6.1 The effect of change-points and non-linear regressors

To investigate the effect of change-points and non-linear regressors, we devised a synthetic toy example, sketched in Figure 13. Consider N=8 random variables, where $X_1,\ldots,X_5$ are iid standard Gaussian distributed. In the first model (Figure 13, left panel), the variables $X_6,\ldots,X_8$ depend on X_5 through a sigmoidal transfer function:

$$X_i = \begin{cases} \epsilon_i & , \ i=1,\ldots,5 \\ \dfrac{2}{1+e^{\frac{-X_5}{\theta}}} + \epsilon_i & , \ i=6,\,7,\,8 \end{cases}. \tag{57}$$

The random noise variables ϵ_i (i=1,…,8) are i.i.d. Gaussian $N(0, \sigma^2)$ distributed. In the second model (Figure 13, right panel), the variables $X_6,\ldots,X_8$ depend on the product term X_4X_5 through a sigmoidal function. For i=6, 7, 8 we have:

$$X_i = \frac{2}{1+e^{\frac{-X_4X_5}{\theta}}} + \epsilon_i \tag{58}$$

where the noise variables ϵ_i (i=6, 7, 8) are i.i.d. Gaussian $N(0, \sigma^2)$ distributed. For overall consistency, all variables were standardized to a standard deviation of 1, and subsequently shifted such that the minimum

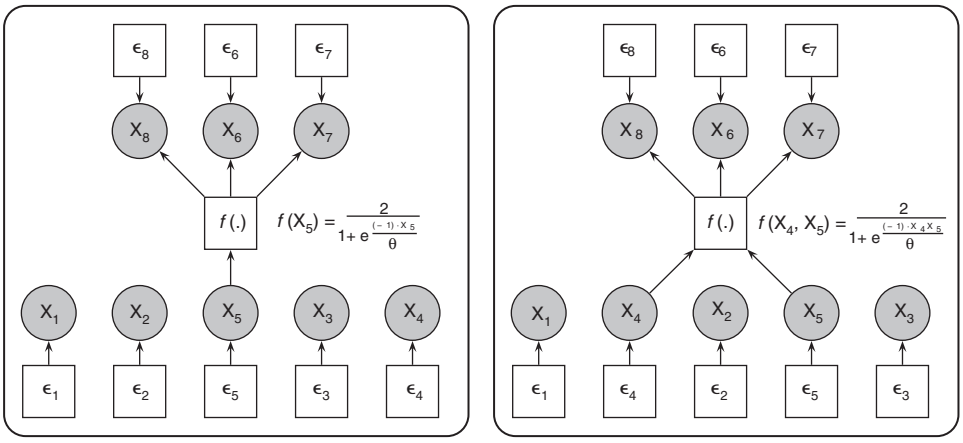

Figure 13 Regulatory network for synthetic data.
The figure shows a graphical representation of the regulatory interactions among the eight variables of the synthetic data described in Section 6.1. In both panels the observed variables, $X_1,\ldots,X_8$, are represented as gray circles, while the (unobserved) random perturbations, $\epsilon_1,\ldots,\epsilon_8$, as well as the non-linear transformation $f(\cdot)$ are represented by white squares. Left panel: The three variables X_6, X_7, and X_8 obtain the same deterministic input, $f(X_5)$, where $f(\cdot)$ is a sigmoidal function. The deterministic signal is perturbed by additive i.i.d. Gaussian noise: ϵ_i (i=6, 7, 8). See main text for further details. Right panel: This graph is similar to the left panel, except that the three response variables X_6, X_7, and X_8 obtain the deterministic input $f(X_4, X_5)$, where $f(\cdot)$ is a sigmoidal function of the product X_4X_5. See main text for further details.

was equal to zero. For both toy scenarios, we generated 25 independent data instantiations with T=100 data points each, from 30 different combinations of the parameters $\sigma^2 \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ and $\theta \in \{1, 0.8, 0.6, 0.4, 0.2\}$.

We first applied the non-homogeneous hierarchical Bayesian regression model from Section 2.6.2 to the synthetic data generated from the model in Figure 13, left panel. The results are shown in the left panel of Figure 14. For low noise levels, $\sigma^2 \leq 0.01$, the network reconstruction accuracy tends to increase with increasing numbers of change-points. Interestingly, the opposite trend is observed for high noise levels, $\sigma^2 \geq 0.1$. This behaviour has the following explanation. A target node, say X_8, depends on the true regressor, X_5, through the non-linear transfer function $f_\theta(\cdot)$ of equation (57); the deviation from linearity increases with decreasing values of θ. On the other hand, there are two covariates, X_6 and X_7, which for low noise levels σ^2 will show a strong linear correlation with the target node X_8. Consider, without loss of generality, node $X_6 = f_\theta(X_5) + \epsilon_6$, which has a linear correlation with the target node, $X_8 = f_\theta(X_5) + \epsilon_8 = X_6 + \epsilon_8 - \epsilon_6 = X_6 + \tilde{\epsilon}$, but subject to double the amount of noise: $\tilde{\epsilon} = \epsilon_8 - \epsilon_6$ implies that $\mathrm{var}(\tilde{\epsilon}) = \mathrm{var}(\epsilon_8) + \mathrm{var}(\epsilon_6) = 2\sigma^2$. Hence, if the transfer function $f_\theta(\cdot)$ is linear, then the true regressor, X_5, is preferred over the spurious one, X_6. However, if the transfer function $f_\theta(\cdot)$ is non-linear, then the model used for network reconstruction needs sufficient non-linear modeling capability to capture the dependence between X_5 and X_8. Otherwise, the spurious variable X_6 will be learned, despite the noise amplification. Now, a non-homogeneous Bayesian regression model with change-points implements effectively a piece-wise linear function and can thus, in principle, approximate the sigmoidal function $f_\theta(\cdot)$. The results depend on the combination of the noise level, σ^2, and the amount of non-linearity, θ. If the noise σ^2 is low, then the effect of the noise amplification, by which the spurious variables are suppressed, is weak, and non-linear modeling capability is critical for good performance, especially as the degree of true underlying non-linearity increases. In that case, more change-points are advantageous and improve the network reconstruction accuracy, as seen from the top rows of Figure 14, left panel. However, piece-wise linear regression models are very flexible and can potentially overfit the data. This tendency towards overfitting gets stronger as the noise level σ^2 increases. In addition, higher noise levels intrinsically suppress spurious variables via the effect of noise amplification, discussed above, thus reducing the need for non-linear modeling capability. As a consequence, more change-points become a disadvantage and deteriorate the network reconstruction accuracy, as seen from the bottom rows of Figure 14, left panel.

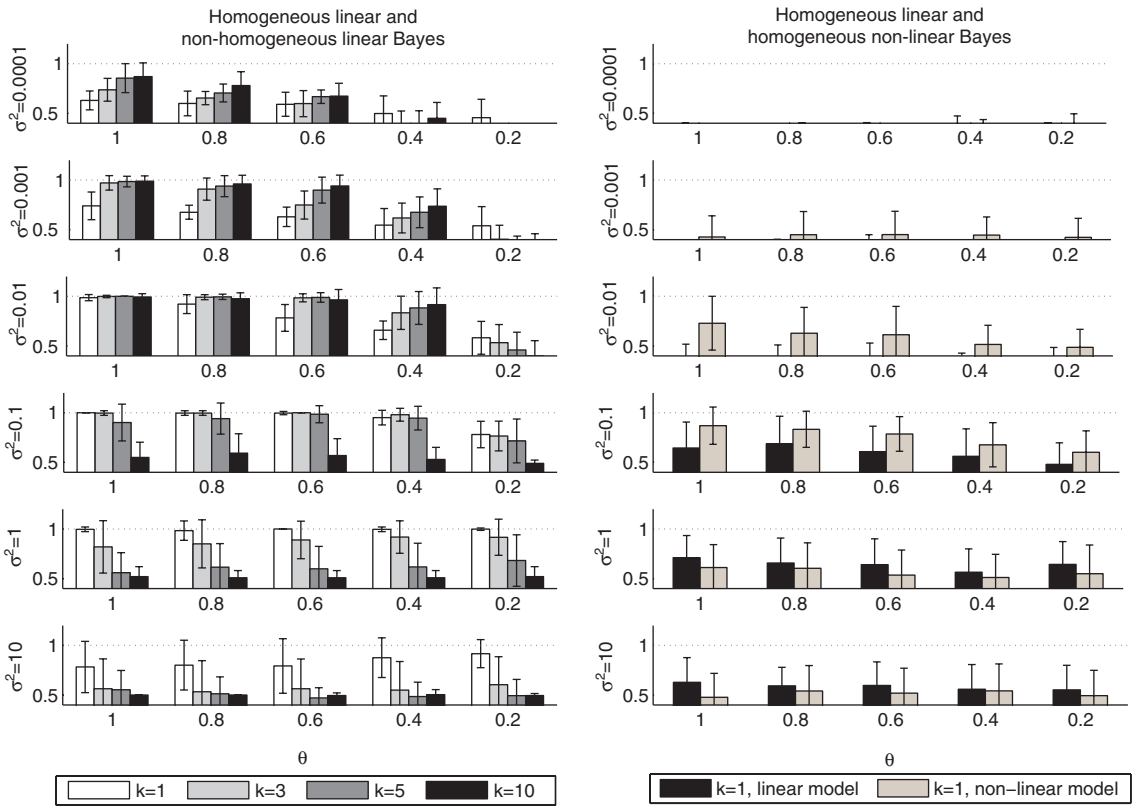

Figure 14 Network reconstruction accuracy on the synthetic data for non-homogeneous and non-linear Bayesian regression models.

Synthetic network data were generated as described in Section 6.1. Left panel: equation (57). Right panel: equation (58). Different parameter combinations of σ^2 (the noise variance) and θ (the interaction strength) were used. The Bayesian regression models described in Sections 2.5–2.6 were applied to network reconstruction. Average AUROC scores were computed from 20 independent data instantiations. Both panels in the figure are arranged as matrices, where the rows correspond to σ^2 and the columns correspond to θ. Left panel: Histograms of the average AUROC scores for the homogeneous Bayesian regression model (white) and three non-homogeneous Bayesian regression models that partition the data with respect to the amplitude of the response variable, with k=3 (light gray), k=5 (gray), and k=10 (dark-gray) segments. The change-point locations were inferred from the data. Right panel: Histograms of the average AUROC scores for homogeneous Bayesian regression models. Black bars refer to the conventional linear Bayesian regression models. Grey bars represent non-linear Bayesian regression models that also include two non-linear transformations of the regressor variables: inverse terms, and quadratic (2nd order) interactions terms. See Section 6.1 for details.

The right panel in Figure 14 compares the AUROC values obtained with two versions of the homogeneous hierarchical Bayesian regression model (Section 2.5): one has only linear terms as regressors (black boxes), the other also includes non-linear (inverse and quadratic) terms (gray boxes). The models were applied to synthetic data generated from the toy network in the right panel of Figure 13. For very low noise levels ($\sigma^2 \leq 0.001$), the network reconstruction is poor. For medium noise levels, ($0.01 \leq \sigma^2 \leq 0.1$), the network reconstruction improves, especially when including non-linear terms as regressors. For high noise levels, the opposite trend is observed: the network reconstruction deteriorates as a consequence of including non-linear terms as regressors. This pattern has a similar explanation as before, following the same trade-off between non-linear modeling capability and noise. A target variable, say X_8, depends non-linearly on two regressors: $X_8 = f_\theta(X_4, X_5) + \epsilon_5$, where $f(\cdot)$ was defined in equation (58). Two confounding covariates, X_6 and X_7, have the same dependence on X_4 and X_5. This leads to a spurious linear association with X_8, subject to noise amplification: $X_8 = X_6 + \epsilon_8 - \epsilon_6 = X_6 + \tilde{\epsilon}$, where $\mathrm{var}(\tilde{\epsilon}) = \mathrm{var}(\epsilon_8) + \mathrm{var}(\epsilon_6) = 2\sigma^2$. For very low noise levels (Figure 14, right panel, top two rows), weakening the spurious linear associations by noise amplification cannot compensate for the approximation errors in modeling the true non-linear interactions; hence the poor performance.

For medium noise levels (Figure 14, right panel, rows 3–4), the spurious linear correlations are suppressed against the non-linear true associations, especially if the model has non-linear approximation power due to the inclusion of non-linear regressors. For high noise levels, (Figure 14, right panel, bottom two rows), noise amplification alone substantially weakens the spurious associations, and additional non-linear modeling capability is counter-productive, due to potential overfitting.

In summary, the upshot of the synthetic toy study is as follows. Even if the true underlying regulatory processes are intrinsically non-linear, additional non-linear modeling capability, in the form of change-points or the explicit inclusion of non-linear terms, is not a panacea for better performance per se. As it turns out, the difference in performance between the linear and non-linear models depends on the amount of intrinsic non-linearity and the noise level. There is a weak trend that a higher degree of intrinsic non-linearity gives the non-linear model an edge on the linear model (Figure 14, left panel, rows 2 and 3). However, a more substantial influence has the noise level. It is only for the lower noise levels that non-linear modeling capability has an advantage. For higher noise levels, it is overshadowed by the susceptibility to overfitting, which leads to a net performance deterioration. This explains why, in Figure 6, the linear Bayesian regression model shows a better performance than its more flexible cousins with change-points or non-linear terms, and why in Figure 11 the performance deteriorates with increasing numbers of change-points. The particular trade-off between non-linear modeling flexibility versus susceptibility to overfitting may vary with the nature of the data generation mechanism, which explains the different trends for mRNAs and proteins in Figure 10.

6.2 The effect of missing protein concentrations

Figure 8 suggests that the network reconstruction accuracy does not improve significantly when including protein concentrations in addition to just mRNA. To understand this counter-intuitive finding, note that two proteins in the circadian clock network, LHY and TOC1, occur in different isoforms, with only one of them acting as transcription factor. If protein data are missing, the gene coding for a regulatory protein has to be taken as a proxy for the regulatory protein itself, both in the modeling as well as in the gold-standard regulatory network. However, the influence of a gene on another gene is indirect, and since both protein isoforms are coded by the same gene, the distinction between isoforms becomes obsolete in the gene regulatory network. If protein data are available, then the model needs to identify the correct protein isoform to obtain a true positive score in the network prediction assessment. Hence, due to the correlation between the concentration profiles of the different isoforms, this is a harder prediction task than the reconstruction of the gene regulatory network from incomplete data (mRNA concentrations only), where this distinction is obsolete. The observation in Figure 8 can thus be explained as a partial compensation of two conflicting tendencies: incomplete data (mRNA only) causes an information loss, which should render the network reconstruction more difficult overall, but it also renders certain aspects of the network reconstruction task easier as a consequence of not having to distinguish between different protein isoforms. The net effect is no significant difference in performance.

6.3 Gaussian process performance

Regarding the poor performance of the GP, we emphasize that we were using the method exactly as described by Äijö and Lähdesmäki (2009), using the authors' own software. The software uses the kernel of equation (33). This is a kernel from the Matérn class, which depends on a further hyperparameter ν; see Rasmussen and Williams (2006) for the explicit expression. The hyperparameter ν defines the degree of roughness, with ν=1/2 giving a rough Ornstein-Uhlenbeck process, and $\nu\to\infty$ reducing to the smooth squared exponential kernel of equation (32). The kernel defined in equation (33) corresponds to ν=3/2.

The GP model from Äijö and Lähdesmäki (2009) thus depends on seven hyperparameters: the mean $\mathbf{b}=(\overline{\alpha},\overline{\lambda})$ and covariance $\sigma_b^2\mathbf{I}$ of the prior distribution on the basal transcription and decay rates, $\boldsymbol{\beta}=(\alpha,\lambda)$,

the Matérn kernel parameters l, a and ν, and the noise variance σ^2. Only three of them are inferred in a maximum likelihood type-II sense: the length scale l, the amplitude a, and the noise variance σ^2. The other four hyperparameters are fixed; these are ν, which defines the roughness of the Matérn class kernel, as well as the parameters that define the prior distribution on the linear parameter vector, $\boldsymbol{\beta}$.

The poor performance of the GP has two possible explanations. First, fixing four of the hyperparameters might be too restrictive, and the chosen Matérn class with ν=3/2 might not be sufficiently representative of the actual concentration profiles. This might indicate that the choice of kernel is quite critical in determining the GP's performance. A second explanation is that for each gene g, the authors choose the set of regulators that maximizes the posterior probability of equation (39). This probability is conditional on the hyperparameters. The methodologically correct approach would be to integrate the hyperparameters out, as e.g., discussed in section 5 of MacKay (1992). The GP method we have applied effectively ignores the last two terms in equation (5.3) of that paper. If the posterior distribution over the hyperparameters is sharply peaked, that will not matter, as integration and optimization then effectively lead to identical results. However, it will make a difference if the posterior distribution is diffuse, in which case the GP model selection is suboptimal. Our results thus suggest that the method presented by Äijö and Lähdesmäki (2009) could be made more powerful with a more rigorous inference scheme, the development of which is beyond the scope of the present paper.

6.4 Comparison with other methods

We briefly discuss the performance of the other methods included in our comparative evaluation. The mutual information based method ARACNE showed the poorest performance. This is not surprising, given that most of the true networks included in our study, shown in Figure 3, violate the premise on which the theoretical foundations of ARACNE are based. The ARACNE network reconstruction theorem states that given some further regularity conditions, ARACNE can correctly reconstruct tree-like networks, i.e., networks containing only pairwise interactions (Margolin et al., 2006). However, there is no theoretical guarantee that densely connected networks or networks containing loops can be correctly reconstructed, and our empirical study suggests that the performance of ARACNE for such networks is, in fact, rather poor.

The poor performance of the Gaussian mixture model (GMM) is presumably due to the fact that model selection is carried out with BIC. BIC is computationally cheap, but over-regularised, leading to structures that are too sparse. Our results are further consistent with the findings by Neuneier et al. (1994) that modeling conditional probabilities indirectly via equation (44) is inferior to modeling them directly with regression-type models, e.g., of the form discussed in Husmeier (1999).

The observation that Tesla shows a slightly poorer performance than Lasso is consistent with the observation that the inclusion of changepoints for the light phases slightly degrades the performance of the hierarchical Bayesian model, as discussed in Section 6.1.

It might be surprising that the Bayesian splines autoregression (BSA) method did not outperform the computationally cheaper linear sparse regression methods Lasso and Elastic Net. This is caused by an oversparsity of the networks predicted with BSA. As discussed by Morrissey et al. (2011), the inclusion of an edge in the network leads to a more substantial increase in the parameter space dimension than for a linear model, due to the fact that an edge is associated with the high-dimensional parameter vector of the splines. Recall from Section 2.8 that the strength of the interaction between two genes g and g', which is modeled with a scalar $w_{g,g'}$ in a linear model, becomes a vector in BSA, $\mathbf{w}_{g,g'}$, spanning the entire range of B-spline basis functions. Hence, the Bayesian approach per se penalises more severely against the inclusion of extra edges than for a linear model, and the non-linear modeling potential of the splines was found to insufficiently compensate for that. We noticed that the performance of BSA improved when the default Jeffreys prior on the edge inclusion probability was replaced by a more informative prior with a concentration of probability mass above 0.5. We have not included these results, because tuning hyperparameters based on the network

reconstruction performance is methodologically incorrect (as it would be using knowledge that is not available in real applications). These findings indicate, though, that the performance of BSA can in principle be boosted by the inclusion of informative prior knowledge. However, even when exploring deviations from the Jeffreys prior, BSA never quite reached the performance of the linear HBR method. This is consistent with the observation that our own non-linear variants of HBR never outperformed the linear version; we refer the reader back to Section 6.1 for a discussion of this trend.

7 Conclusion

We have carried out a comparative evaluation of 15 state-of-the-art statistics/machine learning methods for regulatory network reconstruction, using the central gene regulatory network of the circadian clock in the model plant *A. thaliana* and a series of network modifications. To evaluate the network reconstruction performance objectively from a proper gold standard, we simulated mRNA and protein concentration time series from a published regulatory network structure. The simulations were based on a mathematical description of the individual molecular reactions, modeled with Markov jump processes to capture the intrinsic stochasticity of these events. The data generation process also emulated various experimental interventions carried out in the laboratory, including the knock-out of certain target genes, and the exposure of plants to different artificial light-dark cycles.

In a preliminary study, we have investigated the effects of various model choices and inference settings: the estimation of the optimal regularization parameters in sparse regression (Lasso, Elastic Net, and Tesla), and the choice of both the structure and the parameter priors in hierarchical Bayesian regression. For estimating the regularization parameters, we have shown that cross-validation is slightly preferable to BIC, and that the heuristic modification suggested by Hastie et al. (2001) is counter-productive. For the structure prior of the hierarchical Bayesian regression model, we have found that there is no significant advantage in using a truncated Poisson distribution on the cardinalities of the sets of regulators over a uniform distribution, subject to the same fan-in restriction. For the parameter prior of the hierarchical Bayesian regression model, we have found that the ridge regression prior significantly outperforms the g-prior.

In the main part of our study, we have applied the competing network reconstruction methods to a large variety of data, generated from different network structures, with different status of observation (mRNA only versus mRNA and proteins), and different methods for estimating de novo mRNA transcription rates. We have systematically disentangled the different effects with an ANOVA scheme. Our results confirm various intuitively plausible trends, e.g., that the difficulty of network reconstruction increases with increasing network connectivity, and that for estimating de novo mRNA transcription rates, data smoothing has a beneficial effect. The novel contribution of our study consists in objectively quantifying these effects, in terms of average AUROC score differences associated with the respective main effects in the ANOVA scheme. For the model comparison, we have shown that hierarchical Bayesian regression outperforms all other methods, again objectively quantifying the performance gain.

Our study has also revealed various surprising trends. Since the mechanisms of transcriptional regulation are based on non-linear Michaelis-Menten kinetics, explicitly imbuing the network reconstruction method with non-linear modeling capability via change-points in the response variable or the inclusion of inverse and quadratic terms should generally benefit the network reconstruction performance. Our study has refuted this conjecture. We have carried out further synthetic toy studies to shed light on these effects. Our study suggests that the results vary substantially with the amount of non-linearity and the noise variance, indicating the regimes where explicit non-linear modeling capability is beneficial, or counter-productive.

We have finally applied the best network reconstruction method from the comparative assessment to the mRNA concentration profiles from the TiMet project. The reconstructed network contains several topological features that are consistent with recently published regulatory networks of the circadian clock in *A.*

thaliana. However, the detailed structure clearly differs. This difference is a consequence of the different nature of the methods. For the networks published in the literature, the processes of transcriptional regulation were modeled with ordinary differential equations. The network structures were not selected with rigorous statistical inference; doing that, e.g., with the procedure proposed by Vyshemirsky and Girolami (2008) is computationally prohibitive. The consequence is a considerable degree of reliance on intuition and biological prior knowledge, as evidenced by repeated recent network modifications in the literature (see Figure 12). The methods applied in the present article are based on more abstract models of molecular regulatory interactions, which render objective statistical inference computationally viable. Hence, our understanding of circadian regulation at the molecular level will potentially improve as a consequence of a synthesis of both approaches, which will suggest novel avenues for model adjustment. The proposed network reconstruction methods are particularly useful for linking circadian regulation in plants to metabolism, due to the current absence of detailed hypotheses and reliable mechanistic models.

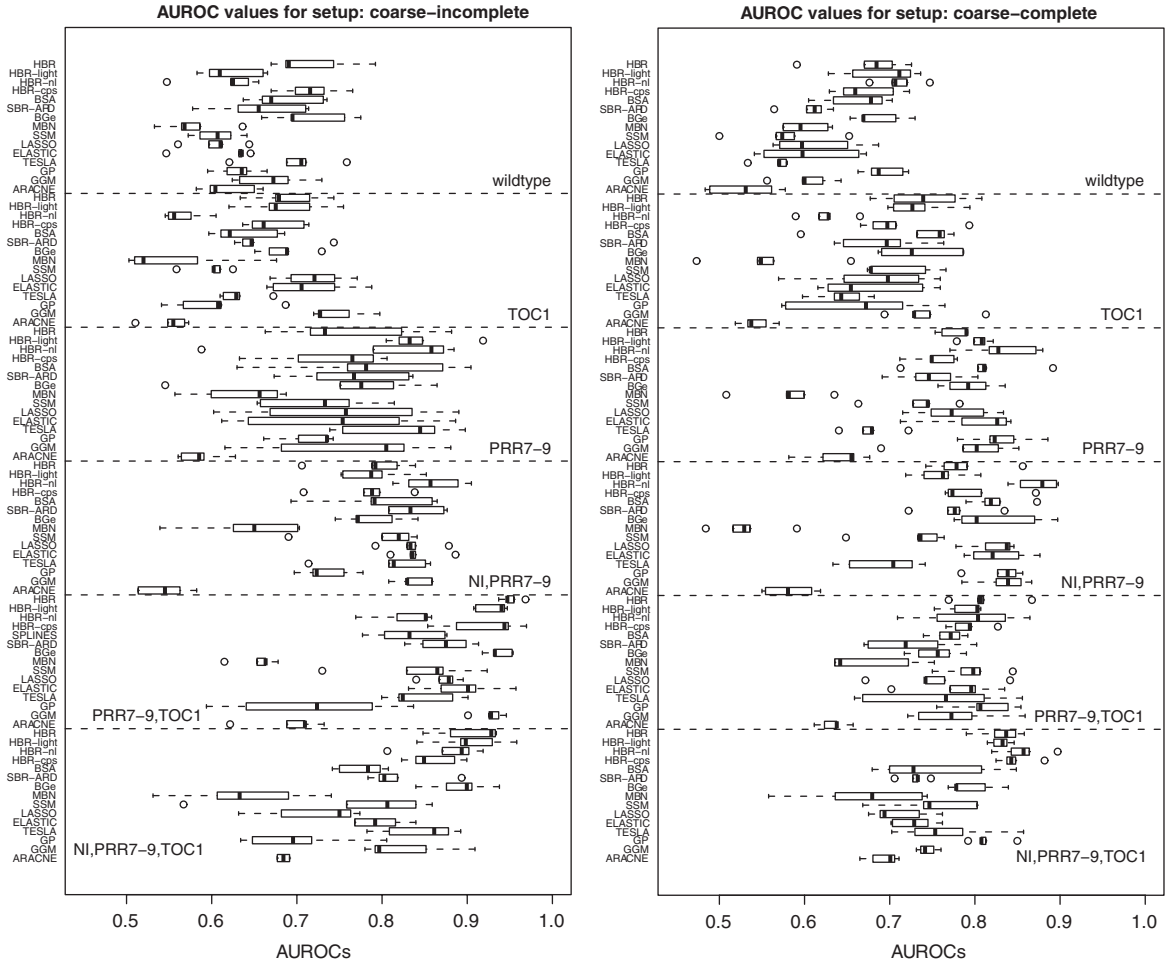

Figure 15 AUROC scores obtained for different reconstruction methods, different network structures, and different experimental settings.

The figures show standard boxplot representations for the distributions of AUROC scores obtained in our study. Only the scores for the coarse response gradients (computed from equation (55) with 2-h intervals) are shown; the corresponding scores for the fine and interpolated gradient are omitted to avoid excessive length of the paper. Left panel: Incomplete data, with mRNA but no protein concentrations. Right panel: Complete data that include both protein and mRNA concentrations. Each panel contains six subpanels, representing the six different network topologies shown in Figure 3.

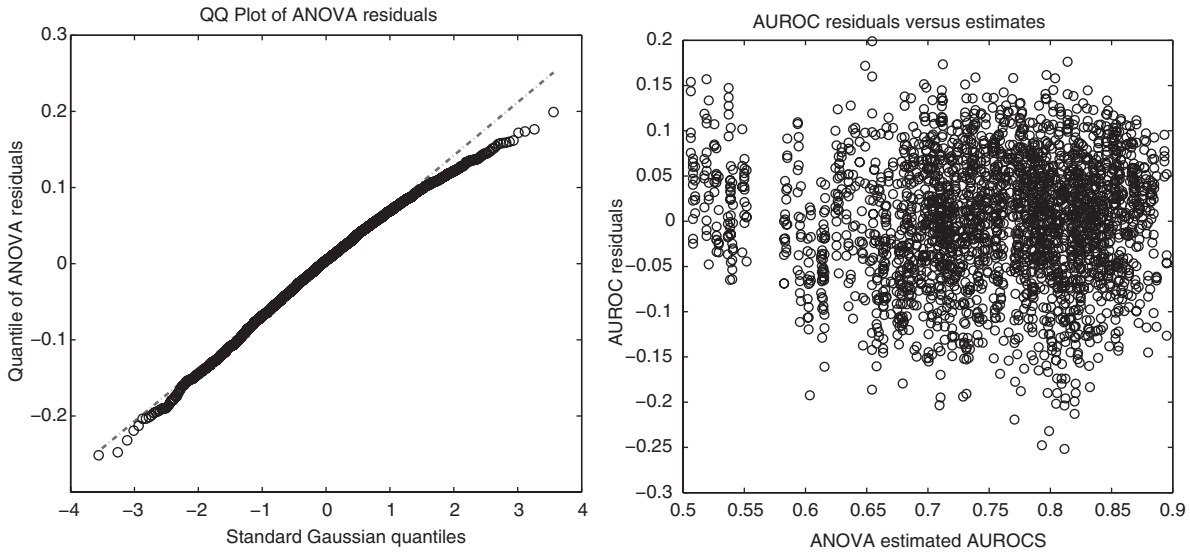

Figure 16 Residual diagnostic for the ANOVA model.
Left panel: QQ-plot. The figure shows a Quantile-Quantile (QQ) plot of the residuals for the ANOVA model, described in Section 4.8, equation (56). The actual quantiles (vertical axis) are plotted against the quantiles of the Gaussian distribution (horizontal axis). The linear relation indicates good agreement with the Gaussian distribution; the deviations for very low and high values point to slightly longer tails. Right panel: Scatter plot diagnostic. The figure shows a scatter plot of the residuals (vertical axis) against the AUROC values fitted with the ANOVA model of Section 4.8, equation (56) (horizontal axis).

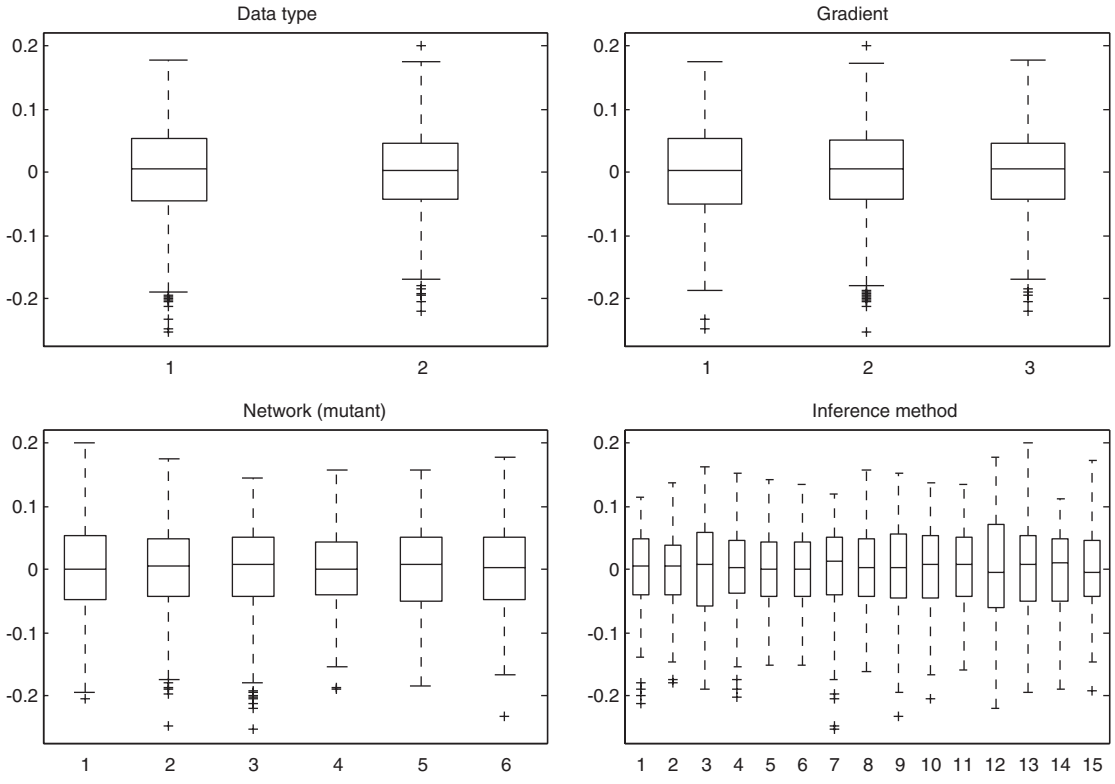

Figure 17 Residual diagnostic for different factors of the ANOVA model.
The figure is arranged as a 2-by-2 matrix, whose four panels correspond to the four main effects of the ANOVA model of Section 4.8, equation (56). Each panel shows a boxplot representation of the distribution of the residuals for all possible values of the corresponding main effects.

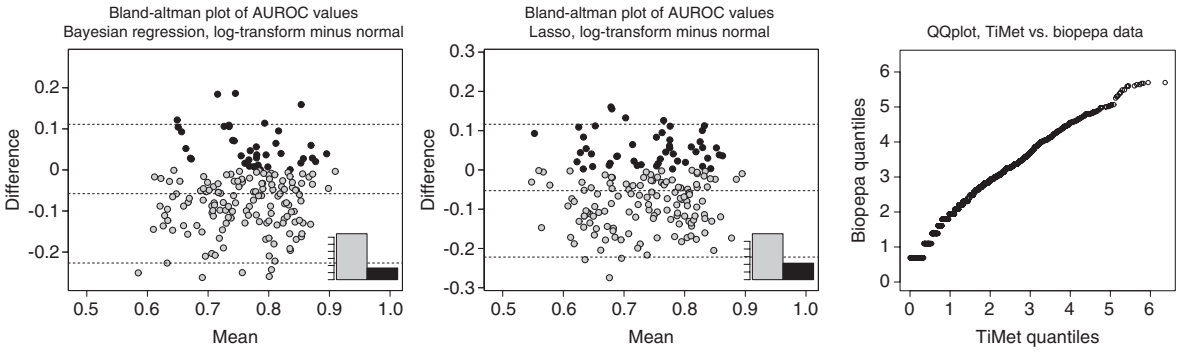

Figure 18 Blant-Altman plot comparing network reconstruction accuracies between log-transformed and original data, and QQ-plot for comparing Biopepa and qRT-PCR data.
The AUROC scores obtained from the original data are compared to those obtained from log-transformed data (the y-axis displays the difference, i.e., log-transformed minus original data). The left panel shows the results when applying the HBR method (avg. difference −0.058) and the center panel for the Lasso method (avg. difference −0.053). For both methods, a majority of negative values can be observed (indicated by the gray box in the embedded histogram), i.e., log-transforming the data is detrimental to the learning accuracy. The right panel displays a QQ-plot comparing the distribution of realistic (Biopepa) and real (qRT-PCR) mRNA concentrations.

Acknowledgments: The work described in the present article is part of the TiMet project on linking the circadian clock to metabolism in plants. TiMet is a collaborative project (Grant Agreement 245143) funded by the European Commission FP7, in response to call FP7-KBBE-2009-3. Parts of the work were done while M.G. was supported by the German Research Foundation (DFG), research grant GR3853/1-1. A.A. is supported by the BBSRC and the TiMet project. We are grateful to Andrew Millar, Alexander Pokhilko, and V. Anne Smith for helpful discussions.

Appendix

A.1 ANOVA for method evaluation

Figure 15 shows some of the raw results from our study. Clear patterns are not immediately discernible by visual inspection, which motivates the ANOVA method described in Section 4.8. To ascertain that the underlying assumptions of the ANOVA model are satisfied, we carried out a standard residual analysis. The objective is to test whether the residuals are independent and identically (i.i.d) normally distributed. A violation of this assumption would indicate that some structure in the data has not been captured by the decomposition of equation (56), and that e.g. higher-order interaction terms would have to be included.

Figure 16, left panel, shows a quantile-quantile (QQ) plot of the residuals to test the assumption of a normal distribution. The straight line confirms that there is good agreement with this assumption overall, with only minor deviations for the lowest and highest quantiles, suggesting that the residual distribution is slightly heavier-tailed.

Figure 16, right panel, shows a scatter plot of all residuals against the corresponding values fitted with the ANOVA model of equation (56). For low values, the spread of the residuals seems to become slightly tighter, but this effect is weak, and overall there is no clearly discernible pattern of any dependence between the residual distribution and the fitted value.

Figure 17 shows histograms of the residuals for all possible values of the four main effects in equation (56). There are no obvious deviations from a uniform pattern, and the results are consistent with the assumption that the distributions of the residuals are identical and independent of the main effects.

These diagnostics thus do not indicate any clear violation of the model assumptions and suggest that the ANOVA model proposed in Section 4.8 provides an adequate mechanism for extracting trends and patterns from our simulations studies.

A.2 Data: comparison between Biopepa and qRT-PCR profiles, and assessing the effect of the log transformation

The right panel of Figure 18 shows a QQ-plot to compare the distribution of mRNA concentrations between the realistic data (Section 3.1) and the qRT-PCR profiles from the Timet project (Section 3.2). There is only a mild deviation from an overall linear dependence, which suggests that the specific technical aspects of qRT-PCR measurements, described, e.g., in Bengtsson et al. (2008), do not require a major modification of our stochastic-process model of transcriptional regulation, as reviewed in Table 3 and implemented in Biopepa. This further suggests that the patterns and trends observed in the comparative evaluation based on our realistic data are indicative of results for real qRT-PCR data, and can be used for providing estimates of expected prediction accuracy and guiding decisions on model choice.

This in particular concerns the decision of whether or not to log-transform the data. Inserting log-transformed concentrations, $\tilde{x}_{g,t}=\log(x_{g,t})$, into the fundamental equation of transcriptional regulation, equation (1), and applying the chain rule of differential calculus yields:

$$\frac{d\tilde{x}_{g,t}}{dt}=[\alpha_g+f_g(\exp(\tilde{\boldsymbol{x}}_{\boldsymbol{\pi}_{g,t}})-\lambda_g\exp(\tilde{x}_{g,t})]\exp(-\tilde{x}_{g,t}) \tag{59}$$

It is seen that in comparison with equation (1), the log transformation has led to a more complicated functional dependence, not only by including an extra multiplicative factor $\exp(-\tilde{x}_{g,t})$ on the right-hand side, but also by making f_g a function of $\exp(\tilde{\mathbf{x}}_{\pi_{g,t}})$, which increases the amount of non-linearity in the system. This suggests that for network reconstruction, a log-transformation of the data will be counterproductive.

To test this conjecture, we have repeated the network reconstruction on the realistic data after subjecting them to a log transformation. The results are summarised in Figure 18, which displays the differences in the form of Blant-Altman plots for the Lasso (center panel) and HBR (left panel) methods. The average AUROC score difference is 0.06 in favor of the original, non-log transformed data. The distribution of paired differences shows that the proportion of negative differences, where the network reconstruction has deteriorated as a consequence of the log transformation, is significantly higher than the proportion of positive differences. This confirms our conjecture that log-transforming the mRNA concentrations is counterproductive. Due to the reasoning in the first paragraph, that patterns observed for the realistic data are indicative of results to be expected for real qRT-PCR data, we have therefore elected *not* to log-transform the Timet data.

References

Ahmed, A. and E. P. Xing (2009): "Recovering time-varying networks of dependencies in social and biological studies," Proc. Natl. Acad. Sci., 106, 11878–11883.

Äijö, T. and H. Lähdesmäki (2009): "Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics," Bioinformatics, 25, 2937–2944.

Andrieu, C. and A. Doucet (1999): "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," IEEE T Signal Proces., 47, 2667–2676.

Barenco, M., D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank (2006): "Ranked prediction of p53 targets using hidden variable dynamic modeling," Genome Biology, 7, R25.

Beal, M., F. Falciani, Z. Ghahramani, C. Rangel, and D. Wild (2005): "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors," Bioinformatics, 21, 349–356.

Beal, M. (2003): Variational Algorithms for Approximate Bayesian Inference, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, UK.

Bengtsson, M., M. Hemberg, P. Rorsman, and A. Ståhlberg (2008): "Quantification of mRNA in single cells and modeling of RT-qPCR induced noise," BMC Molecular Biology, 9, 63.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*, Singapore: Springer.

Brandt, S. (1999): *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, New York, USA: Springer.

Brooks, S. and A. Gelman (1999): "General methods for monitoring convergence of iterative simulations," J. Comput. Graph. Stat., 7, 434–455.

Butte, A. J. and I. S. Kohane (2000): "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing*, volume 5, 418–429.

Ciocchetta, F. and J. Hillston (2009): "Bio-PEPA: A framework for the modeling and analysis of biological systems," Theor. Comput. Sci., 410, 3065–3084.

Davies, J. and M. Goadrich (2006): "The relationship between Precision-Recall and ROC curves," Proceedings of the 23rd International Conference on Machine Learning, 233–240.

Edwards, K., O. Akman, K. Knox, P. Lumsden, A. Thomson, P. Brown, A. Pokhilko, L. Kozma-Bognar, F. Nagy, D. Rand, A. J. Millar. (2010): "Quantitative analysis of regulatory flexibility under changing environmental conditions," Mol. Syst. Biol., 6, 424.

Feugier, F. and A. Satake (2012): "Dynamical feedback between circadian clock and sucrose availability explains adaptive response of starch metabolism to various photoperiods," Front. Plant Sci., 3.

Friedman, J., T. Hastie, and R. Tibshirani (2008): "Sparse inverse covariance estimation with the graphical Lasso," Biostatistics, 9, 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," J. Stat. Softw., 33, 1–22.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000): "Using Bayesian networks to analyze expression data," J. Comput. Biol., 7, 601–620.

Geiger, D. and D. Heckerman (1994): "Learning gaussian networks," in International Conference on Uncertainty in Artificial Intelligence, Seattle, WA: Morgan Kaufmann Publishers, 235–243.

Gelman, A. and D. Rubin (1992): "Inference from iterative simulation using multiple sequences," Stat. Sci., 7, 457–472.

Gillespie, D. (1977): "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem., 81, 2340–2361.

Grzegorczyk, M. and D. Husmeier (2012): "A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology," Stat. Appl. Genet. Mol. Biol. (SAGMB), 11, article 7.

Grzegorczyk, M. and D. Husmeier (2013): "Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models," Mach. Learn., 91, 1–50.

Guerriero, M., A. Pokhilko, A. Fernández, K. Halliday, A. Millar, and J. Hillston (2012): "Stochastic properties of the plant circadian clock," J. R. Soc. Interface, 9, 744–756.

Hanley, J. A. and B. J. McNeil (1982): "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, 143, 29–36.

Hastie, T., R. Tibshirani, and J. J. H. Friedman (2001): The Elements of Statistical Learning, volume 1, New York: Springer.

Herrero, E., E. Kolmos, N. Bujdoso, Y. Yuan, M. Wang, M. C. Berns, H. Uhlworm, G. Coupland, R. Saini, M. Jaskolski, A. Webb, J. Gonçalves, S. J. Davis. (2012): "EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock," Plant Cell, 24, 428–443.

Husmeier, D. (1999): Neural Networks for Conditional Probability Estimation: Forecasting Beyond Point Predictions, Perspectives in Neural Computing, London: Springer.

Husmeier, D. (2003): "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," Bioinformatics, 19, 2271–2282.

Kalaitzis, A. A., A. Honkela, P. Gao, and N. D. Lawrence (2013): *gptk: Gaussian processes tool-kit*, URL http://CRAN.R-project.org/package=gptk, R package version 1.06.

Ko, Y., C. Zhai, and S. Rodriguez-Zas (2007): "Inference of gene pathways using Gaussian mixture models," in International Conference on Bioinformatics and Biomedicine, Fremont, CA, 362–367.

Ko, Y., C. Zhai, and S. Rodriguez-Zas (2009): "Inference of gene pathways using mixture Bayesian networks," BMC Syst. Biol., 3, 54.

Kolmos, E., M. Nowak, M. Werner, K. Fischer, G. Schwarz, S. Mathews, H. Schoof, F. Nagy, J. M. Bujnicki, and S. J. Davis (2009): "Integrating ELF4 into the circadian system through combined structural and functional studies," HFSP J, 3, 350–366.

Lawrence, N. D., M. Girolami, M. Rattray, and G. Sanguinetti (2010): Learning and inference in computational systems biology, Cambridge, MA: MIT Press Cambridge.

Lèbre, S., J. Becq, F. Devaux, G. Lelandais, and M. Stumpf (2010): "Statistical inference of the time-varying structure of gene-regulation networks," BMC Syst. Biol., 4.

Locke, J. C. W., M. M. Southern, L. Kozma-Bognár, V. Hibberd, P. E. Brown, M. S. Turner, and A. J. Millar (2005): "Extension of a genetic network model by iterative experimentation and mathematical analysis," Mol. Syst. Biol., 1.

Locke, J. C. W., L. Kozma-Bognár, P. D. Gould, B. Fehér, E. Kevei, F. Nagy, M. S. Turner, A. Hall, and A. J. Millar (2006): "Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*," Mol. Syst. Biol., 2.

MacKay, D. J. (1992): "Bayesian interpolation," Neural Comput., 4, 415–447.
Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano (2006): "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," BMC Bioinformatics, 7.
Marin, J.-M. and C. P. Robert (2007): Bayesian core: A practical approach to computational Bayesian statistics, New York, USA: Springer.
Meyer, P. E., F. Lafitte, and G. Bontempi (2008): "minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information," BMC Bioinformatics, 9.
Morrissey, E. R., M. A. Juárez, K. J. Denby, and N. J. Burroughs (2011): "Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression," Biostatistics, 12, 682–694.
Murphy, K. P. (2012): Machine learning: a probabilistic perspective, Cambridge, MA: MIT Press.
Nabney, I. (2002): NETLAB: algorithms for pattern recognition, Springer.
Neuneier, R., F. Hergert, W. Finnoff, and D. Ormoneit (1994): "Estimation of conditional densities: a comparison of neural network approaches," in *International Conference on Artificial Neural Networks*, National Cheng Kung University, Taiwan: Springer, 689–692.
Opgen-Rhein, R. and K. Strimmer (2007): "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data," BMC Syst. Biol., 1.
Pokhilko, A., A. Fernández, K. Edwards, M. Southern, K. Halliday, and A. Millar (2012): "The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops," Mol. Syst. Biol., 8, 574.
Pokhilko, A., S. Hodge, K. Stratford, K. Knox, K. Edwards, A. Thomson, T. Mizuno, and A. Millar (2010): "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," Mol. Syst. Biol., 6.
Pokhilko, A., P. Mas, A. J. Millar, et al. (2013): "Modeling the widespread effects of TOC1 signaling on the plant circadian clock and its outputs," BMC Syst. Biol., 7, 1–12.
Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani (1996): "The DELVE manual," URL http://www.cs.toronto.edu/delve.
Rasmussen, C. E. (1996): Evaluation of Gaussian processes and other methods for non-linear regression, Ph.D. thesis, Citeseer.
Rasmussen, C. and C. Williams (2006): Gaussian processes for machine learning, volume 1, MA: MIT press Cambridge.
Rogers, S. and M. Girolami (2005): "A Bayesian regression approach to the inference of regulatory networks from gene expression data," Bioinformatics, 21, 3131–3137.
Schäfer, J. and K. Strimmer (2005): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," Stat. Appl. Genet. Mol. Biol., 4.
Smith, M. and R. Kohn (1996): "Nonparametric regression using Bayesian variable selection," J Econometrics, 75, 317–343.
Solak, E., R. Murray-Smith, W. E. Leithead, D. J. Leith, and C. E. Rasmussen (2002): "Derivative observations in Gaussian process models of dynamic systems," Advances in Neural Information Processing Systems, MIT Press: Vancouver, Canada, 1033–1040.
Tibshirani, R. (1995): "Regression shrinkage and selection via the Lasso," J. R. Stat. Soc. Series B, 58, 267–288.
TiMet (2014): "The TiMet Project - Linking the clock to metabolism: URL http://timing-metabolism.eu.
Tipping, M. and A. Faul (2003): "Fast marginal likelihood maximisation for sparse Bayesian models," in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL, 1, 3–6.
Tipping, M. (2001): "Spare Bayesian learning and the relevance vector machine," Journal of Machine Learning Research, 1, 211–244.
Vyshemirsky, V. and M. Girolami (2008): "Bayesian ranking of biochemical system models," Bioinformatics, 24, 833–839.
Weirauch, M. T., A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, DREAM5 Consortium, Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolvitzky, G, and T. R. Hughes (2013): "Evaluation of methods for modeling transcription factor sequence specificity," Nat. Biotechnol., 31, 126–134.
Werhli, A. V., M. Grzegorczyk, and D. Husmeier (2006): "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks," Bioinformatics, 22, 2523–2531.
Wilkinson, D. J. (2009): "Stochastic modeling for quantitative description of heterogeneous biological systems," Nat. Rev. Genet., 10, 122–133.
Wilkinson, D. (2011): Stochastic modeling for systems biology, volume 44, Taylor & Francis, Boca Raton, FL: CRC press.
Zoppoli, P., S. Morganella, and M. Ceccarelli (2010): "TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by ab information theoretic approach," BMC Bioinformatics, 11.
Zou, H. and T. Hastie (2005): "Regularization and variable selection via the Elastic Net," J. R. Stat. Soc. Series B, 67, 301–320.